Time-Series Embedded Feature Selection Using Deep Learning: Data Mining Electronic Health Records for Novel Biomarkers

Gavin Tsang

Submitted to Swansea University in fulfilment
of the requirements for the Degree of Doctor of Philosophy

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

September 19, 2022

Declaration

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ..... (candidate)

Date .....

Statement 2

I hereby give my consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

Abstract

As health information technologies continue to advance, routine collection and digitisation of patient health records in the form of electronic health records present as an ideal opportunity for data-mining and exploratory analysis of biomarkers and risk factors indicative of a potentially diverse domain of patient outcomes. Patient records have continually become more widely available through various initiatives enabling open access whilst maintaining critical patient privacy. In spite of such progress, health records remain not widely adopted within the current clinical statistical analysis domain due to challenging issues derived from such "big data".

Deep learning based temporal modelling approaches present an ideal solution to health record challenges through automated self-optimisation of representation learning, able to manageably compose the high-dimensional domain of patient records into data representations able to model complex data associations. Such representations can serve to condense and reduce dimensionality to emphasise feature sparsity and importance through novel embedded feature selection approaches. Accordingly, application towards patient records enable complex modelling and analysis of the full domain of clinical features to select biomarkers of predictive relevance.

Firstly, we propose a novel entropy regularised neural network ensemble able to highlight risk factors associated with hospitalisation risk of individuals with dementia. The application of which, was able to reduce a large domain of unique medical events to a small set of relevant risk factors able to maintain hospitalisation discrimination.

Following on, we continue our work on ensemble architecture approaches with a novel cascading LSTM ensembles to predict severe sepsis onset within critical patients in an ICU critical care centre. We demonstrate state-of-the-art performance capabilities able to outperform that of current related literature.

Finally, we propose a novel embedded feature selection application dubbed 1D convolution feature selection using sparsity regularisation. Said methodology was evaluated on both domains of dementia and sepsis prediction objectives to highlight model capability and gener-

alisability. We further report a selection of potential biomarkers for the aforementioned case study objectives highlighting clinical relevance and potential novelty value for future clinical analysis.

Accordingly, we demonstrate the effective capability of embedded feature selection approaches through the application of temporal based deep learning architectures in the discovery of effective biomarkers across a variety of challenging clinical applications.

Acknowledgements

During my time within Swansea University studying for my Ph.D., extending through the "unprecedented times" of a global pandemic and the various other ups and downs experienced within a Ph.D. candidature; I have the pleasure to have known, worked with, and been supported by a wide variety of incredible individuals across the years. To thank every one of you would require another thesis length document by itself. Nevertheless, I extend my sincerest gratitude and best wishes to everyone, without which, I would not be in the position I am now.

In particular, I wish to give thanks to the following individuals for going above and beyond in supporting me in this endeavour. Professor Xianghua Xie, my primary supervisor, for guiding and supporting me throughout the years from being a clueless graduate student to a slightly less clueless Doctor of Philosophy. Xianghua has always been understanding and considerate of my academic and personal life and has driven me towards this great achievement. This thesis would not be possible without Xianghua and as such, I am eternally grateful.

I also wish to extend my gratitude towards my second supervisor Dr. ShangMing Zhou in addition to the Swansea Vision research group and co.: Dr. Mike Edwards, Dr. Jingjing Deng, Dr. David George, Dr. Joss Whittle, Dr. Alex Lee, Michael Kenning, and the many others for their knowledge and friendship for when we discussed research or vented frustration.

I would also like to thank my family, Chei Mei Tsang, Yu Tao Tsang, and Kelly Tsang who, without their loving care and support, I would not have been able to have achieved so much. Finally, my appreciation also goes to my wonderful friends, in particular Nicole Almond and Lucy Zou, for continually being a source of laughter and joy throughout my time in university.

Again, I wish express my appreciation to everyone for their contribution, big or small, to this incredible achievement. Thank you.

Contents

List of Tables

List of Figures

List of Symbols

Chapter 3: Machine Learning

Symbol	Definition
w	Weight coefficient vector for SVM.
$\phi(x)$	Mapping function for SVM.
b	Bias coefficient for SVM.
L	Model loss dictated by distance error between prediction and label.
x	Input feature.
y	Class label.
J	Number of features in the model/layer input vector.
a	Optimal Lagrange multipliers for SVM.
$K(x_i, x_j)$	Kernel mapping matrix for SVM.
G	Gini impurity, dictating probability of incorrect classification.
c	Index of class label.
C	Number of class labels in a dataset.
$p(c)$	Probability of class, c.
A	Covariance matrix of a dataset, X.
V	Matrix of eigenvectors.
Λ	Diagonal matrix of eigenvalues.
Σ_B^-	Between class distance for LDA.
Σ_W^-	Within class distance for LDA.
x_i	Input feature value.
$p(x\|C_k)$	Probability of feature vector x, given class C_k.
l	Index of current model layer in a neural network.
k	Index of feature activation output from the current network layer, l.
K	Number of features in the current network layer, $l-$.
a_k^l	Activation output of a perceptron indexed as k at layer l.
w_{jk}^l	Weight coefficient of corresponding link between features, j and k at layer l.
b_k^l	Bias coefficient of the corresponding perceptron indexed as k, at layer l.
$\sigma()$	Activation function for perceptron.
y_i	Class label for input, i.
$\hat{y}_i$	Model prediction for input, i.
W^l	Weight coefficient matrix for layer l.
δ^L	Model overall distance error.
δ^l	Layer, l, distance error.
η	Learning rate used in stochastic gradient descent.
t	Time-step within a timeline.
C_t	Cell state of an LSTM at timestep, t.
f_t	Forget gate vector of an LSTM.
i_t	Input gate vector of an LSTM.
C_t'	Encoded input vector to be combined to create cell state, C_t.
h_t	Output activation of a LSTM cell at current time-step, t.
W_O	Weight coefficient matrix of the output gate of an LSTM.
b_O	Bias coefficient vector of the output gate of an LSTM.
$MI(x, y)$	Mutual information between input, x, and output, y.
$H()$	Information entropy.
Ω	Generic regularization function.
λ	User defined weighting hyper-parameter for a generic regularization function.
L_1	L_1 normalization.
L_2	L_2 normalization.

Chapter 5: Applied EHR Modelling

Symbol	Definition
x	Generic input value.
$P(x)$	Probability mass function dictating information entropy.
$L(w)$	Loss penalty dictated by weighted information entropy.
λ	User defined weighting hyper-parameter emphasising entropy regularization.
l	Index of current model layer in a neural network.
j	Index of feature activation output from the previous network layer, $l-1$.
J	Number of features in the previous network layer, $l-1$.
k	Index of feature activation output from the current network layer, l.
K	Number of features in the current network layer, $l-$.
w_{jk}	Weight coefficient of corresponding link between features, j and k.
t	Index of training epoch.
T	Number of training epochs used in model optimization.
M	Number of snapshot sub-models to be produced.
$\alpha(t)$	Learning rate indicated at training epoch, t.
α_0	Learning rate at beginning of training (epoch 0).
R_k	Feature importance value of feature k, which in this case would be the input layer.
W_k	Vector of weights all associated to feature, k.
σ^2	Function producing variance of a vector.

Chapter 6: Dementia Hospitalisation

Symbol	Definition
m	Index of sub-model produced by boosted cascading methodology.
M	Number of sub-models dictated by user.
t	Time-step within a patient timeline.
T_i	Number of time-steps within a specific patient, i
i	Index of patient timeline.
L^m	Loss induced by sub-model m within the forward propagation training procedure.
y_i^t	Class label of a time-step, t, for patient, i.
$\hat{y}_i^{t,m}$	Model, m, prediction of a time-step, t, for patient, i.
w_i^m	Weighting factor coefficient dictating patient, i, importance used for sub-model, m
λ_w	User defined weighting hyper-parameter emphasising boosting strategy over cascading strategy.
β	Weight coefficient of hinge loss function.
β_0	Bias coefficient of hinge loss function.
ξ_i	Margin of error allowing for non-perfect prediction of patient, i, in a hinge loss function.
C	User defined hyper-parameter constant dictating proportional distance error allowed by predictions.
L_C	Loss penalty dictated by the critical diagnosis point penalty function.
t_{sepsis}	Time-step of initial indication of sepsis within a patient time-line.
t_{opt}	Time-step of optimal prediction point, prior to t_{sepsis}.
t_{late}	Time-step of latest positive prediction point of sepsis before significant penalty is induced.
λ_C	User defined weighting hyper-parameter emphasising critical diagnosis point penalty function.
$\lambda_T P$	User defined weighting hyper-parameter emphasising weighting of true positive predictions in critical diagnosis point penalty function.
λ_e	User defined weighting hyper-parameter emphasising weighting of early positive predictions in critical diagnosis point penalty function.
L_N	Loss penalty dictated by the negative reversal penalty function.
λ_N	User defined weighting hyper-parameter emphasising negative reversal penalty function.

Chapter 7: Linear Aggregation Kernel Based Feature Ranking

Symbol	Definition
t	Time-step within a patient timeline.
x^t	Input feature vector of a patient at time-step t.
i	Index of input feature in input vector x.
N	Number of features within input feature vector x_t.
y^t	Class label feature vector of a patient at time-step t.
k	Index of linear aggregation kernel.
K	Number of linear aggregation kernels, k, used in model.
x'_k	Resulting embedded activation output of linear aggregation kernel, k.
$w_{i,k}$	Weight coefficient of the corresponding feature input, x^i linked to kernel, k.
$L(w)$	Loss penalty function, sparse regularization, dictated by weight, w.
λ_1	User defined hyperparameter used in sparse regularization. This dictates maximum loss penalty of the function.
λ_2	User defined hyperparameter used in sparse regularization. This defines penalty curve of the function.

Chapter 1

Introduction

Contents

1.1 Motivations

Within the current information age, digital information technologies permeate every facet of human life allowing for unprecedented interconnected communication of information spanning the globe. Consequentially, the concept of big data—an increasingly understated title by modern trends[1]—generates immense hype and investment in every field of science and industry[2, 3]. For modern society, big data heralds new levels of multi-disciplinary scientific discovery and economic value, promising discovery and analysis of large scale population trends and heterogeneities never before possible with small-scale data. For the data scientist, big data presents significant and unique computational and statistical challenges including scalability, data noise, spurious correlations, incidental endogeneity, and measurement errors[4] with non-trivial applications and solutions.

Machine learning (ML) based approaches, inductive generalisation of data, allows for the production of automated predictive modelling and discovery applications; an ideal solution for the automated processing of excessive amounts of data into actionable information, placing value on data. Statistical pattern recognition methods such as the humble naive Bayes (NB) linear regressor and support vector machines (SVMs), to ensemble weak learner approaches such as random forests (RFs) and neural networks (NNs) have been adapted to numerous and diverse ranges of applications over the decades. Recent state-of-the-art in ML focuses upon deep learning (DL) approaches[5], coinciding with the big data based advances in digital technologies regarding computational capability and data capacity. The increasingly mainstream deep neural network (DNN) family of DL approaches have maintained state-of-the-art capabilities, able to self-optimize feature representations based off of observed data. As such, the DNN family have seen significant application in numerous big data research domains such as computer vision[6], natural language processing[7, 8], and medical decision support systems. The latter domain of which, health informatics, is of particular emphasis. Examples of successful application include studies highlighting the potential health outcome benefits of ML based approaches applied to real-world clinical environments[9] whilst a large proportion of studies report significant improvements to diagnosis and prognosis prediction across a wide variety of medical applications from neuroimaging[10, 11, 12] to health records[13, 14, 15] to genome modelling[16, 17, 18].

1.1.1 The Electronic Health Record

The domain of health information technology has seen great advancement with big data based technology improvements. The collection and storage of patient health information in the form of electronic health records (EHRs) provide insight into individualised health across a diverse population. Modern advances in information linkage[19] enables extensive patient follow-up across the many disparate medical systems, enabling EHR timelines to extend to potentially entire lifetimes; whilst anonymisation technologies[20] enable such data to be widely accessible in research contexts without risks towards data privacy exposure.

As such, the application of state-of-the-art DL based big data approaches on the relatively untapped domain of EHRs[21, 22, 23] promises great advancement in the concepts of individualised medicine and improved patient outcomes[24, 25].

However, with such rich information potential, the application of the EHR in current clinical statistical analysis for implementation of improved care delivery is highly limited[21, 22, 23]. Traditional statistical modelling, prevalent within the evidence-based medicine approach practised on small sample size clinical trials, remains highly limited when applied to big data based EHRs. Consequently, within EHR based health informatics we identify two key domains, of which jointly, there exists potential advancement in novel approaches to the analysis of individualised health information.

Temporal Information

Of significance is the temporal nature of patient health, continually changing from age and underlying conditions. Such longitudinal information is critical in capturing the underlying trends and thus complete health picture of an individual as time progresses. Within traditional statistical model-driven health informatics literature, longitudinal health analysis focus on cross-sectional type approaches[26]. Applications of such relying on simultaneous modelling of correlation between all temporal observations through autocorrelation regression estimation[27] or via moving-average autoregressive models[27]. Cross-sectional type approaches remain popular due to the alternative consideration of the complete time-series, a challenging approach due to the continual increase in feature-set size with the inclusion of a temporal dimension. The utilisation of the complete temporal dimension can be achieved through methodologies such as recurrent ML-based modelling approaches which will be of focus within this thesis.

Feature Dimensionality

Large data dimensionality is one such characteristic of big data. EHRs present a prime example of significantly high dimensionality through the sheer potential of unique medical events experienced by the individual patient across a lifetime. In conjunction to the big data challenges mentioned previously, rules based medicine approaches rely on the concept of simplicity and human comprehension. Consequentially, the application of high dimensional data is itself undesirable whilst being infeasible for the medical professional to evaluate and apply to a decision support system. The required submission of a full patient history remains impractical for the diagnosis of an individual whilst remaining impossible to validate due to the intractable, black-box nature of such a modelling approach. Current literature has emphasised hand-selection of small feature size EHRs[28] reducing the potential capability of big data based large-scale population analysis. Such approaches remain limiting for exploratory based studies such as discovery of novel biomarkers due to the prior assumption of relevance or lack-thereof in the set of considered features. As such, approaches with consideration for novel exploratory biomarker discovery presents considerable potential with which this thesis will bring to focus.

The two concepts of temporal information and feature dimensionality remain critical to effective exploitation of the EHR to it's full potential as a big data based domain. Said concepts however, remain relatively unexplored for the former, or run counter to current established medical approaches in the latter case. Consequently, we seek to incorporate the domain of big data based DL approaches towards producing novel and effective solutions to our established challenging use case.

1.1.2 Objective

Within a big data domain such as EHRs, there generally exists a large feature redundancy and subsequently small relevance for a significant proportion of medical events towards the stated modelling objective or patient outcome. Furthermore, the application of EHR analysis for a clinical based objective requires significant limitations in dimensionality to ensure feasibility as a practical tool. Within current literature, such feature selection has been based on *a priori* expert or domain knowledge with little consideration for exploratory analysis of the wider set of potential features available within a big data based EHR.

There exists an entire domain of ML focused feature selection approaches, categorised as filter, wrapper and embedded methodologies; said research domain of feature selection will be

explored in detail in chapter 3. The focus of this thesis will be on embedded type methodologies; in particular, we seek to apply DL based architectures as the modelling and prediction component of our EHR based application and to leverage the self-optimised feature representation capabilities as the foundation for novel feature selection approaches. The use of DL architectures enables the utilisation of the temporal specific, recurrent based sub-category of DL architectures to exploit the longitudinal progression of the individual's health.

The primary objective of this thesis is accordingly, the exploration and development of novel approaches to biomarker discovery within a clinical objective or condition through EHR data-mining. Development approaches will focus on embedded type feature selection on DL based modelling methodologies, specifically recurrent based DNNs in order to exploit self-optimisation of feature representation within a temporal context as the selection criteria foundation. Such approaches remain young in maturity within the larger domain of ML and to a greater extent within the health informatics domain of EHR analysis. As such, there exists potential for novel and opportune research approaches within said domain.

In conjunction, the secondary objective of this thesis concerns validation of our proposed feature selection approaches on sufficiently complex and relevant clinical challenges existing within a real population dataset as a case study. The development of a clinically feasible application for the prediction of said patient outcomes would enable potential significant improvements to individualised medical care through optimisation and direction of clinical utilization towards critical patients. Of greater significance is the discovery of relevant and potentially novel biomarkers, indicators of a particular patient outcome through which, greater understanding can be driven towards through traditional clinical research trials. We seek to evaluate such biomarkers for clinical relevance as a whole, to ensure validity of our proposed methodologies.

The case studies in question will be pertaining to sepsis development within a critical care context and dementia based hospitalisation event risk within a long-term analysis of an individual; clinical relevance and motivations of said objectives are detailed in greater detail in chapter 4. Whilst the focus will remain solely on said cases within this thesis, the proposed methods will remain wholly generalisable to a large range of clinical applications. Such generalisability is due to the open, unfocused nature of EHRs detailing a complete patient health picture with which, our proposed methodologies data-mine. Method generalisability is also heuristically proven through validation with the use of two distinctly unique-in-characteristic case studies and the application of a large variety of unique EHR datasets.

1.2 Contributions

The main contributions of this thesis can be seen as follows:

- A comprehensive review of EHR based machine learning applications within the current state-of-the-art literature in the context of dementia risk and of sepsis development. Through which we present current relevance of said topics within the greater medical domain, the challenges of applying ML based approaches to such a clinical objective, the current state-of-the-art methodologies and studies within said domain and finally future research pathways and opportunities for further study, of which we approach in the following contributions.

- A novel approach to feature selection within a use-case study, identifying at-risk individuals with dementia in encountering a hospitalization event. The proposed methodology consists of an ensemble architecture of DNNs trained using the "snapshot ensemble" approach to aid in reducing over-fitting and perturb feature weighting in combination with novel entropy weight regularisation to produce sparse feature representations. Such representations are thus applied as ranking and selection criteria to produce a final selected feature-set.

- A novel architecture and training approach towards alleviating issues of high-dimensionality, data sparsity, and class imbalance to produce a prediction tool for sepsis development able to outperform the current state-of-the-art approaches within literature according to the PhysioNet 2019 CinC Challenge. The proposed approach consists of a continuation of the ensemble approach detailed previously, incorporating a novel boosted cascading architecture approach.

- A novel embedded feature selection methodology incorporating linear aggregation kernels in combination with long short-term memory (LSTM) recurrent networks and a novel weight sparsity regularization approach to produce sparse feature representations within the aggregation kernel. Through which, we extrapolate feature importance, or lack-thereof, to produce a final set of highly relevant features. The proposed methodology was validated through case study using both clinical objectives of predicting sepsis development and separately, dementia development across two uniquely characterised EHRs.

- We present a complete list of discovered biomarkers found to be of relevance towards the proposed case studies. Through which, we analyse the clinical relevance and novelty value to highlight potential avenues of novel clinical research and validation. Said list highlights a significant proportion of known risk factors relevant towards our case studies, highlighting model selection capability whilst several novel medical events were highlighted with minor to no studies exploring such correlation.

Outcomes from this thesis have also contributed towards several publications across a range of conferences and journals. Following on is a list of published and in review publications.

- **Mining Electronic Health Records to Identify Influential Predictors Associated with Hospital Admission of Patients with Dementia: An Artificial Intelligence Approach**
  Zhou, S.M., Tsang, G., Xie, X., Huo, L., Brophy, S., and Lyons, R.A
  2018 The Lancet 392, S9

- **Harnessing the Power of Machine Learning in Dementia Informatics Research: Issues, Opportunities, and Challenges**
  Tsang, G., Xie, X., and Zhou, S.M.
  2019 IEEE Reviews in Biomedical Engineering 13, 113-129

- **Modelling Large Sparse Data for Feature Selection: Hospital Admission Predictions of Dementia Patients using Primary Care Electronic Health Records**
  Tsang, G., Zhou, S.M., and Xie, X.
  2020 IEEE Journal of Translational Engineering in Health and Medicine 9, 1-13

- **Deep Learning Based Sepsis Intervention: The Modelling and Prediction of Severe Sepsis Onset**
  Tsang, G., and Xie, X.
  2020 25th International Conference on Pattern Recognition, 8671-8678

- **Convolution Based Feature Ranking: Identifying Predictive Indicators within Electronic Health Records**
  Tsang, G., and Xie, X.
  2021 Journal of Biomedical and Health Informatics **[UNDER REVIEW]**

1.3 Outline

The remainder of the thesis is arranged and outlined as follows:

Chapter 2 Electronic health records

This chapter presents background on the EHR, detailing current motivations, challenges and opportunities in the utilisation of EHRs within ML based applications.

Chapter 3 Machine Learning

This chapter introduces the larger domain of ML, emphasis is placed on DL and time-series based ML approaches. Following on is an introduction to current state-of-the-art in feature selection approaches currently applied within the context of EHRs.

Chapter 4 Applied EHR Modelling: An Overview of Clinical Objectives

This chapter highlights the background motivations, challenges and current state-of-the-art in analysis and prediction of our chosen case studies: dementia based risk analysis and sepsis development.

Chapter 5 Dementia Hospitalisation: Risk Factor Identification Using Entropy Cascades

In this chapter we present our first proposed methodology in feature selection, the entropy cascade NN. Experimental evaluation is performed by case study of dementia based prediction of hospitalisation risk.

Chapter 6 Modelling Severe Sepsis Onset: Boosted Cascading LSTMs

The focus of this chapter is in modelling approaches towards alleviating the big data challenges of EHRs. The boosted cascading LSTM methodology is proposed within this chapter and evaluated by sepsis onset prediction, presenting state-of-the-art performance capabilities.

Chapter 7 Linear Aggregation Based Feature Ranking: Identifying Predictive Indicators within Electronic Health Records

We present within this chapter, the linear aggregation kernel based feature ranking methodology for feature selection. We evaluate on both sepsis and dementia case studies. Through which, we present impressive feature selection capability able to select a small subset of medical events able to maintain effective modelling performance whilst being clinically relevant.

Chapter 8 Conclusions and Future Work

We finally draw concluding remarks and retrospective on the presented chapters and novel applications. We aim to highlight potential drawbacks, assumptions and challenges encountered, looking to inform future potential avenues for continued development of our presented work.

Chapter 2

Electronic Health Records

Contents

2.1 Introduction

With the modern age of health information technology (HIT), routine collection and digitization of patient healthcare data, ranging from patient histories to medication to demographics, produce extensive collections of individualized medical information across a diverse and wide-ranging population which exist in the form of electronic health records (EHRs). Even in consideration of the many disassociated systems of distinct health-care providers within a region; in combination with modern capabilities for patient data linkage[29], there exists an unprecedented and complete longitudinal perspective of an individual's health records from birth to death. EHRs provide substantial opportunity for large-scale, exploratory analysis via modern information technology approaches in the hopes of increasing understanding and improving individualized patient care outcomes and care utilization[21, 22].

Despite the rich information potential of EHRs, it is still not widely adopted in current clinical statistical analysis for use in improved care delivery[21, 22, 23]. Such lack of EHR adoption remains hampered by concerns of data quality and validation, completeness and heterogeneity between disparate data systems globally[30]. Not least, the development of common traditional clinical statistical models suffers the non-trivial obstacle of scalability on the continually expanding domain of EHRs: the challenges of big data and the "curse of dimensionality".

Presented in this chapter is a broad overview of EHRs and it's place in medical informatics, machine learning based modelling approaches. We first introduce the concept of EHR data in section 2.2 for understanding. Following on, is a discussion of the potential opportunities and current landscape of EHR application within the medical informatics field in section 2.3. This leads us onto a overview in section 2.4 of the current challenges in EHRs within a machine learning context to highlight the non-trivial context of such an application.

2.2 Electronic Health Record Collection and Content

With the primary purpose of the EHR being documentation of patient histories and treatment for reimbursement, such accumulated individualized patient information provides a highly desirable platform for analysis. To this end, the diverse approaches between healthcare providers in structuring EHRs to facilitate ease of record keeping reflects poorly on ease of use in health informatics analysis. As such, the formatting of health information can be categorized, in order of convenience, into: structured, semi-structured, and unstructured data.

2.2.1 Structured Data

In the context of medical informatics, highly structured data, stored within rigid fixed schema database tables, presents a straightforward approach to analysis. Examples of such commonly being: demographic information (e.g. birth date, race, gender); event codes (e.g. diagnostic, procedure and medication codes); healthcare encounter information (e.g. admission and discharge); and vital signs (e.g. blood pressure, pulse, weight, height). Structured data presents straightforward analysis through significant standardization across disparate healthcare providers. An example of which being standardized medical encoding systems such as the International Classification of Diseases (ICD) system or the UK National Health Service (NHS) read code system.

Solely structured data however, remains impractical and thus infrequent in real-world EHR systems; requiring anticipation and codification of every possible data element at every event entry, rendering such a system unfeasible and unusable through overly high complexity. Consequently, information must be stored and conveyed in an approach not conducive to the rigid requirements of an entirely structured data system.

2.2.2 Unstructured Data

Free text, unstructured data represents the opposite extreme of the structured data concept. Said data consists of narrative text–which includes general practice (GP) encounter notes, specialist reports, admission-discharge summaries, etc.–resulting in highly qualitative, diverse, and detailed information not possible within the rigid confines of structured data. With such rich and diverse information potential, the organization into a form applicable for analysis however, is a non-trivial task.

Narrative text as an avenue for analysis suffers greatly from inconsistent structure or framework providing highly varied and subjective health information. Such variance extends past care providers of varying specialities, to institutions with differing or unique policies, to nationalities with unique dialect or language challenges. The use of natural language processing (NLP) tools to preprocess and extract relevant knowledge from free text represents a non-trivial task, presenting challenges inherent of language comprehension requiring state-of-the-art NLP applications[8].

2.2.3 Semi-structured Data

Semi-structured data forms the majority of EHR structures consisting of flexible name-value-timestamp triplets representing a wide range of medical events based upon the indicated name field from laboratory measurements to dispensed medication to procedure and diagnostic codes associated with an individual patient. An example of such being an "arterial blood pressure" measurement of 145mmHg performed at 16:38, 20/12/21. Expandability is afforded through the definition of newly named events without requiring the restructuring of database tables. Event indications through name and timestamp entry further ensures database capacity remains manageable and non-sparse by eliminating irrelevant event non-occurrence entries and irrelevant periods of time apparent within a structured, rigidly defined schema of set events and regular timesteps of a solely structured data concept. Consequently, semi-structured systems remain the most common approach for recording EHRs.

Semi-structured data, whilst benefiting as the median of the data structure scale, provide unique informatics challenges due to the name-value-timestamp triplet design. Akin to issues with unstructured free text EHRs, dynamic indications of events by name and timestamp present highly varying and inconsistent health pictures of patients dependant on the subjective views of the care provider. Without consistent or regular recordings of pre-set critical health information, two patients suffering of the same condition can present highly differing and consequently informatively sparse health pictures unconducive to consistent traditional machine learning (ML) based modelling approaches. Data challenges also extend towards collation of several multi-institution, nationwide EHR systems with inconsistent naming or codification of events across care-provider systems. Such issues are alleviated through aforementioned standardised coding systems such as ICD and NHS read codes. However, even consistent use of a set coding system in dataset comparisons only generally apply at a national level and not world-wide thus requiring a system of normalising and mapping across systems.

2.3 Opportunities

Modern medical understanding and practice is based upon the foundation of evidence-based medicine[31]. The principles of which, produce large collections of evidence-based best-practice recommendations suggested by committee with validation through randomized clinical trials. Said randomized clinical trial is considered the gold standard research vehicle for validation and evidence creation. The clinical trial however, suffers from small population

sample sizes resulting in issues of population bias and uncertainty in validity for the larger diverse global population. Larger sample sizes and longer-term clinical trials allow for more certainty in validation however; said clinical trial results in significant time commitments and high expense to conduct[22]. Consequently, the clinical trial, whilst essential, is relegated to a confirmatory role in *a priori* hypotheses with little room for exploration based approaches.

EHR based data mining and analysis approaches through long-term observational studies provide a complimentary solution for exploratory analysis through the creation of a large variety of novel possible hypotheses able to be validated through the clinical trial. With large scale, potentially nation-wide population coverage and individual follow-up durations extending from potentially birth to death, EHRs sit at a uniquely optimal position for extensive exploratory analysis. Additional adoption of automated modern ML and health informatics based applications further enhances potential with extensive big data data mining and analysis tools ideal for EHRs.

Whilst the field of ML based EHR health informatics is still young; continued maturing has led to several recent reviews of literature within areas relating to ML, EHRs, and feature ranking highlight prevailing trends, limitations, and possible avenues[21, 28, 32]. Consequently, there exists a large selection of related review and study literature exploring and analysing the various avenues and opportunities available. Following on, we will highlight the foundation of several avenues of research and understanding in which EHRs have or will provide novel effective solutions.

2.3.1 Cohort Selection

A complementary aspect for clinical trials, cohort selection was traditionally performed via manual patient chart reviews to identify eligibility for study inclusion. Large scale EHRs able to record longitudinal individualised information (with emphasis on demographics, medical history, and lifestyle) enable greater categorization and fine identification of ideal cohort candidates. Whilst large general coverage of a population expands quantity and diversity of available candidates lessening population bias concerns.

Such processes of distinguishing patients on patient records accordingly become extremely time-consuming and challenging depending on criteria complexity. Automation of cohort selection is generally achieved through phenotyping applications, via hand-crafted rules based selection or ML selection algorithms. Phenotype definitions involve physical or biochemical traits such as specific diseases or conditions, physical characteristics or lifestyle choice. As in-

dicated by literature review, most common phenotypes of interest are identified as cancer and diabetes comprising the significant proportion of included studies[33]. Various initiatives exist which seek to provide and automate phenotyping applications for cohort selection purposes. Examples of such national initiatives being the UK Biobank[34] for the United Kingdom or the Electronic Medical Records and Genomics (eMERGE) Network for the United States.

2.3.2 Medical Trajectory & Patient Outcomes

Studying the historical, natural progression of a disease or condition is an obvious question of great importance in medical care. Again, the large scale and longitudinal nature of EHRs present an abundance of instances of diseases and said progression over long periods of time. Of importance, associated medical trajectories—initiated by the same condition and culminating in the same patient outcome—can diverge and vary highly between individuals from healthy to condition development to associated complications[35]. Consequently, said application of large scale individualised records for analysis and study can lead to the development of personalised, tailored treatments aimed at improving individual patient outcomes: the modern concept of personalised healthcare[36, 37].

Associated with medical trajectory is the study, measurement and prediction of patient outcomes. As noted previously, the modern trend of healthcare is shifting towards the concept of personalised healthcare[36], consequently maximising the "value" of healthcare services on the individual patient measured by patient outcome compared to the relative cost[38]. Patient prognosis, otherwise the prediction of risk or probability of patient outcome, is a foundational aspect of health care[39]. Outcomes are often defined as specific events such as hospital discharge or death, or measured by quantities such as disease progression or quality of life. The prediction of which is shaped by individual circumstances and health conditions. Such predictions aim to inform healthcare providers on severity and thus adjust intensity of required management, with a proven record of success[24, 25].

The ubiquity of EHRs, and subsequent patient information capability, introduces a new avenue in producing increasingly more advanced, automated risk prediction tools through ML based applications. There exists a large collection of studies applying said concept to fields such as cardiovascular disease risk prediction[25, 40] and dementia risk prediction[41]. Whilst there has been considerable success in producing risk prediction models able to outperform traditional manually measured clinical risk prediction tools, there still exists many challenges including improving clinical understanding with "black box" ML models and application gen-

eralisability concerns progressing from varying individual perspectives to hospital systems to the diversity of international healthcare systems.

2.3.3 Epidemiology & Biomarker Discovery

Epidemiology: studying the determinants, dynamics and distribution of diseases within populations has been of consistent interest within the medical field; significantly, the advent of the SARS-CoV-2 (COVID-19) pandemic has contributed to rising attention in the public eye of the importance of effective epidemiological study.

Biomarker discovery, the study of identifying key indicators contributing to the development of a condition or disease, runs parallel to risk prediction tools previously highlighted. Of significance, effective risk prediction runs secondary to the identification and understanding of the underlying risk factors. In particular—aforementioned "black box" ML based risk prediction models, able to produce state-of-the-art results whilst being unable to illustrate the underlying approach or risk factors, remains a significant drawback of practical implementation raising concerns such as guaranteed generalisability without the aforementioned true understanding of a model to manually ensure rationality. Biomarkers provide a succinct summary of an individual patient's health in relation to the medical condition in question. Similarly, risk factors provide potential to highlight at risk individuals applicable for intervention style treatment.

Data mined risk factor identification remains a highly under-utilized methodology in conjunction with ML based risk prediction applications. Adversely, biomarker selection within ML based modelling studies remains rooted in prior domain/expert knowledge with small subsets of known features[28]; subsequently limiting the full capability of big-data ML algorithms and feature-rich EHR patient information.

Of importance, risk factor discovery or biomarker discovery within EHRs through ML based methodologies, is highlighted as the primary objective focus of novel contributions within this thesis. As demonstrated in future chapters, the data mining of feature-rich EHR information provides substantial potential in automation and identification of current and novel biomarkers. In regards to current risk factor identification ML applications in use with EHRs within literature, generalized linear models are the most common algorithms within the literature whilst regularized regression methodologies generally incorporated feature selection via stepwise approaches[28]. Recent literature reviews have indicated deep learning (DL) based feature selection as an underdeveloped possible avenue, warranting further exploration[42],

especially in the domain of EHR applications[43]. Of the supervised, non-DL methodologies employed recently, most feature selection methods can be categorised into filter and wrapper methods each with respective advantages and disadvantages[42, 44] whilst embedded methods are comparatively less popular[42]. Of the existing embedded feature selection methods, algorithms such as support vector machine with recursive feature elimination (SVM-RFE) and neural network (NN) pruning are most commonly utilized[44].

2.3.4 Multi-morbidity & Adverse Event Reporting

With continuing advancements in healthcare and subsequent life expectancy, the proportion of elderly population continues to increase globally[45]. Of significance, age represents a significant risk factor of increasing multi-morbidity and polypharmacy—resulting in an increasingly elderly population with multiple chronic health disorders and thus the usage of multiple medications for treatment respectively. The consequence of both aspects, resulting in significant concerns of lower quality of life and worse health outcomes[46]; exponential expenditure on healthcare costs as multi-morbidity increases[47]; and unanticipated adverse side effects through multi-disease and multi-drug interaction[48]. Such concerns have lead to calls towards improving chronic disease management, disease interaction knowledge, and combined multi-morbidity treatment plans[49]. Significantly, development of computerised decision support systems to aid in personalising treatment guidelines to the individual multi-morbidity case; of which EHRs and ML play a significant role.

There exists several significant developments in regard to multi-morbidity and polypharmacy in relation to EHR systems. Advantages of global connectivity and collection of health information result in national open-access, self-report initiatives such as the UK Yellow Card Scheme (YCS)[50] and the US FDA Adverse Event Reporting System (FAERS)[51] pharmacovigilance systems to name a few. Initiatives such as YCS and FAERS provide potential for analysis application such as causality assessment[52] or adverse event prediction[53].

As shown, there exists numerous opportunistic avenues of research direction available. Within said potential avenues, the domain of research of this thesis emphasises patient outcome prediction in conjunction with biomarker discovery. As mentioned, EHRs remain a relatively untapped resource of significant information potential. Such information potential is however hampered by non-trivial challenges limiting traditional, statistical analysis techniques. An overview of said challenges will be examined in detail following on.

2.4 Challenges

EHR based data-mining presents significant opportunity and potential in improving individualised patient healthcare as previously mentioned. Said application of EHRs however, pose a variety of non-trivial challenges both unique and inherent to individualised human health information in the form of EHR data and due to being such big-data applications respectively. Said challenges will be briefly explored within this section.

2.4.1 "Big Data"

Inherently, "big data" aspects of EHR data results in similar issues prevalent within other related data applications—namely: high dimensionality, sparsity, noise, and complex non-linear relationships and dependencies between data features. Further challenges of being time-series based long-term longitudinal data compound said big data difficulties significantly. Such challenges, whilst common to ML applications within other applied fields, present non-trivial difficulties which require to be addressed through methodological solutions or domain knowledge incorporation. For instance, prior studies have surprisingly been highly limited in predictor utilization in comparison to the high-dimensional potential of EHRs. Recent systematic reviews of the literature highlight the median count on variable use at close to only 30 variables[28, 32]. A stark contrast to examples such as the common ICD system, codifying and hierarchically classifying 14,000 possible unique medical terms[54]. The adoption of longitudinal information, a major strength of EHR data, is also severely under-utilized within recent literature[28]; instead relying on consolidated patient information and traditional non-temporal statistical analysis[22, 28, 43].

2.4.2 Data Privacy

The implementation of EHRs present significant and complex patient privacy concerns in research[21, 28]. Continued progress in open access availability through information governance policies such as patient de-identification, has opened avenues in EHRs based research[55, 32]. Digital initiatives exist which bring EHRs towards enabling research-based, online, open-access to real-world patient records whilst maintaining ethical patient privacy and national data protection laws; examples of such being Secure Anonymised Information Linkage (SAIL)[56] or Medical Information Mart for Intensive Care dataset (MIMIC)[57] to name a few with relevance towards this thesis.

2.4.3 Data Linkage

Data fragmentation is a prevalent challenge within EHR systems due to division, at all levels, of the many components of healthcare into disparate specialist care providers, departments, hospitals, and national systems; each only having access to fragmented and duplicated partial information. Such continuously expanding, large and expansive (potentially) national data systems requires automated, near-perfect data integration solutions to ensure reliability for research applications; a highly complex and non-trivial task which falls under the field of data linkage[19, 20]. The implications of imperfect data linkage is discontinuity in EHR data either from partial missing data due to a lost linkage or incorrect data from wrongly linked patient records. Both of which produce potentially significant cohort bias and noise layered on top of already non-trivial naturally occurring data challenges highlighted within this section.

2.4.4 Observational Analysis

EHR data mining is a purely retrospective, observational analysis of historical patient records; consequently, EHRs suffer from similar drawbacks of non-controlled experimental environments available to clinical controlled trials. Aspects such as follow-up appointment and data-collection frequency become inconsistent patient-to-patient, resulting in highly irregular time-series data. Compounded with already considerable data sparsity and potentially considerable periods of time between relevant information entries leads to significant difficulty. Adversely, collection frequency varies highly depending on information; patient information such as vital signs are routinely collected as opposed to events such as laboratory tests required only as relevant to patient diagnosis and treatment plans. The analysis of regular time-series data is a consistently well studied research application in ML, however application of such techniques to EHR irregular time-series data is highly challenging[22].

Further related challenges in observational EHR analysis involve effective cohort selection. Studies performed using EHRs often have population biases[58] due to lack of control over cohort characteristics and unknown confounding variables. Such aspects can be moderately controlled through careful selection of control and positive case cohorts to ensure balanced cohort demographics and characteristics. Adversely, studied conditions and diseases can be presented with significantly low prevalence rates throughout an observed population; the result of which is highly skewed cohort size proportions towards the larger control class. Class imbalance is further exacerbated by the aforementioned culling of unsuitable patients to ensure unbiased cohorts. Significant reductions in cohort sizes to small subsets runs in contrary to

the big data challenges of high dimensionality and sparsity requiring substantial cohort sample sizes to ensure ML modelling challenges such as over-fitting do not occur. Such opposing aspects produce a delicate balance requiring non-trivial and novel approaches to ensure acceptable performance and validation.

2.4.5 Missing & Incorrect Data

Missing or incorrect data with EHR patient records is a consistent issue with subsequent consequences of population biases and confounding effects. In elaboration—omitted data, such as critical relevant diagnosis events, produce patient samples of potentially directly wrong classification or indirectly produce large variation within features due to missing critical predictors in samples; incorrect data produces similar issues. Missing and incorrect data is symptom to a variety of data management and entry or physical real-world issues which leads to significant compounding difficulties on top of natural population bias and confounding effects.

Aforementioned data handling issues include omitted or modified medical events due to medical or financial reimbursement policies arising case-by-case within billing records. Consequently, EHRs originating from billing record systems do not necessarily represent a true picture of an individuals medical record. Additional data handling issues include revisions to code definitions, the addition or significant modification of event codes or categories, obfuscating historical events without researcher consideration for a defined mapping between previous and current versions. Recent examples include a new revision to the commonly used ICD coding system from version 10 to 11 with the addition of traditional medicine, sexual health, and new addiction events included[59]. Incorrect data entry mistakes remain a continuous issue across all fields including medical informatics which involve database systems.

Real-world complications focus on more human aspects which result in incorrect data and missingness. Of note, coinciding challenges also inherent to the clinical trial and treatment plans is patient compliance and retention. The latter being issues such as loss to follow-up, patients missing dictated follow-up appointments or a lack thereof post-conclusion of a treatment plan; the consequence of a highly positive patient outcome without lasting issues. As such, there exists biases towards negative patient outcomes due to those being primarily recorded within EHRs whilst positive outcomes are largely missing but assumed. Loss to follow-up is also an issue caused by potential data linkage complications. An unobservable source of missing data is patient compliance with dictated prescriptions, lifestyle changes, and other such interventions; an age-old healthcare challenge[60]. For example, medication prescriptions make

up a significant proportion of EHR data for the individual patient but have no guarantee to prove patients actually take the medication. Interventions requiring greater patient commitments such as lifestyle changes result in lower compliance rates[61].

Chapter 3

Machine Learning

Contents

3.1 Introduction

The capability of machine learning (ML) has continued to expand significantly over recent decades following the expansion of micro-processing power and electronic storage density. ML—inductive generalisation of data, producing automated predictive modelling and discovery applications—presents an ideal solution for the automated processing of data into actionable information. ML methodologies have expanded into all facets of data analysis applications from high-speed, automated stock trading applications within the financial system[62, 63]; to computer vision methodologies identifying brain degradation in magnetic resonance imaging (MRI) scans attributed to dementia[64].

Traditional ML methodologies are commonly reliant on significant data pre-processing to ensure accurate modelling. For instance: high data variance and bias[65], feature multi-collinearity[66], high dimensionality and sparsity[5], and non-linearity are examples of data properties which highly influence ML capability, requiring pre-processing or *a-priori* consideration.

The ubiquity of large-scale personal information records on all facets of life, a consequence of the popularity of the internet and internet of things (IoT) systems[67], presents a significant challenge to condense such large-scale "big-data" into comprehensible and actionable information to create value[68]. Electronic health record (EHR) represents one such domain within big-data requiring non-trivial solutions to the issue of big-data, amongst other unique challenges. Of significance, the complex challenges attributed to EHRs (e.g. cohort bias, confounding variables and multi-collinearity, high dimensionality and sparsity, complex non-linearity) present significant difficulties to traditional ML modelling techniques. EHR and it's challenges are presented in more detail within chapter 2. Consequently, medical informatics applications with EHRs generally fall into traditional large pre-processing pipelines or the more recent application of deep learning (DL).

This chapter presents an overview into the various ML and DL methodologies currently in use within the field of EHR based research. Initially, there will be a quick overview of traditional ML models still in use within current literature followed by an in-depth view of DL based applications, current state-of-the-art, and its relevance towards this thesis. Finally there will be a discussion on feature selection processes in conjunction with ML applications.

3.2 Traditional Modelling Methodologies

3.2.1 Support Vector Machine

Support vector machine (SVM) is a supervised learning model which attempts to generate separating hyperplanes across groups of observations in accordance with class labels. First proposed by Cortes *et al.* [69] in 1995. Unlike linear discriminant analysis (LDA), SVM makes no assumptions on data distribution, allowing for great flexibility in model generation.

By mapping observations from original feature space into a higher dimensional space through the use of linear or non-linear kernel functions, observations which were once non-linearly separable in feature space may be mapped into a higher dimensional space which supports separation by linear hyperplanes.

The separating hyperplanes within higher dimensional space are defined by

$$w^T \phi(x) + b = 0 \tag{3.1}$$

where w is the normalised normal vector to the hyperplane, b the normalised perpendicular distance of the hyperplane to the origin and $\phi(x)$ the linear or non-linear mapping function. The resulting classification function for any new observations is simply comparing the observation position in relation to constructed hyperplane eq. (3.1). Since there exists an infinite set of hyperplanes which could potentially separate class boundaries, an optimal separating hyperplane must be generated based upon the structural risk minimisation principle. As such, the optimal separating hyperplane is arranged so that there is the greatest separating distance, or *margin*, between the borders of class distributions defined as parallel hyperplanes. The superficial observations which lie on the aforementioned parallel hyperplanes are called support vectors.

The optimal separating hyperplane can be found by maximising said margin distance using the Lagrangian dual of the optimisation function

$$\arg_a \min L_p(a) = \sum_{i=1}^{l} \sum_{j=1}^{l} a_i a_j y_i y_j K(x_i, x_j) \tag{3.2}$$

subject to constraints:

$$\sum_{i=1}^{l} a_i y_i = 0 \tag{3.3}$$

$$a_i \geq 0 \text{ for } i = 1, \ldots, l \tag{3.4}$$

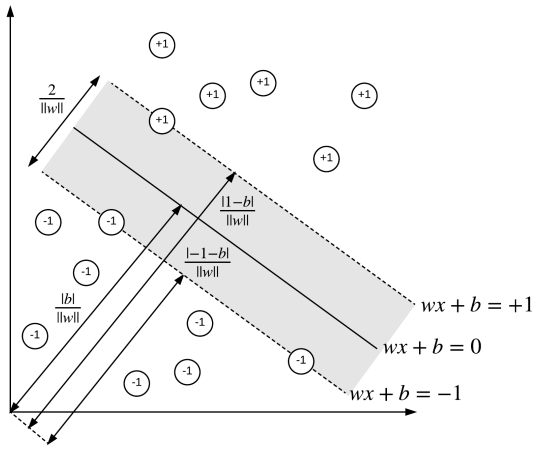

Figure 3.1: Diagram of SVM separating hyperplane across observations of a binary class $\{-1, +1\}$. The optimal separating hyperplane produces the maximum margin distance between class boundaries defined as parallel hyperplanes which lie on superficially located observations called support vectors.

where a are optimal Lagrange multipliers found through quadratic optimisation and $K(x_i, x_j)$ is the kernel mapping matrix.

3.2.2 Random Forest

Random forests (RFs) are an ensemble learning method involving the generation of multiple decision trees whose input dataset are a random sample of features with replacement (*feature bagging*) and also a random sample of observations with replacement (*tree bagging*)[70]. Typically in a classification problem of J features, a subset of $\lfloor \sqrt{J} \rfloor$ features are selected[71] for each tree. Final classification involves the aggregation, generally vote count, of prediction result from every tree.

The use of ensemble classification in RF reduces overfitting whilst also allowing for the evaluation of feature importance after training using the out-of-bag (OOB) error[71]. By permuting individual features across a dataset passed into a trained RF for evaluation, the resulting OOB error can be compared against the original training OOB error to determine feature importance with a greater difference indicating a greater importance for said feature and vice versa. The OOB error serves as a validation metric without the need for a entirely separate validation dataset by evaluating prediction accuracy of observations on decision trees who's training subset did not contain said observations.

Decision trees involve the generation of a directed acyclic graph in a tree like structure

containing interior nodes corresponding to individual features containing edges whose conditional response is based on the set of possible values. Leaf nodes would correspond to the final classification or regression result. Feature selection of interior nodes is evaluated based on various possible metrics such as the traditional Gini impurity G, dictating probability of incorrect classification within a sample sub-set split:

$$G = \sum_{c=1}^{C} p(c)(1 - p(c)), \tag{3.5}$$

where C is the number of classes and $p(c)$, the probability of class c within the chosen subset split. Subsequently, evaluation of split effectiveness is the weighted sum Gini impurity across both sides of the split, where weighting is determined by proportional sample size within each sub-set split. Optimization is thus, the maximisation of ΔG between previous subset and optimised newly split sub-set.

3.2.3 Principal Component Analysis

Principal component analysis (PCA), proposed by Pearson[72], is a methodology used to orthogonally transform a set of observations containing potentially correlated features into a set of linearly uncorrelated features called principal components. Principal components indicate the vector dictating the direction of greatest variance in a normally distributed dataset, *eigenvector*, whilst the *eigenvalue* corresponds to the variance along said vector. Subsequent principal components provide the next greatest variance along the orthogonal vector of all preceding eigenvectors.

Through the eigen-decomposition of the covariance matrix, A, of a dataset x, given by $A = x^T x$, A can be decomposed into the matrix of eigenvectors V where $V_{i,:} = (v_1, \ldots, v_k)$ and eigenvalues Λ

$$A = V \Lambda V^T \tag{3.6}$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$. For each eigenvalue λ_i, a specific eigenvalue equation exists which can be solved for to determine the set of eigenvectors associated to each eigenvalue. Whilst solving for such equations are trivial on small datasets, large feature and observation datasets are solved through the use of various iterative algorithms.

A common use of PCA involves its use as a precursor to dimensionality reduction. Lower order principal components of low variance or eigenvalue can be removed while higher order principal components are kept, reducing dimensionality whilst retaining as much variance in

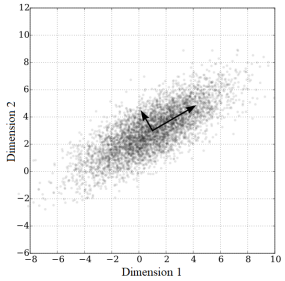

Figure 3.2: PCA of a gaussian distribution showing the orthogonal eigenvectors. The arrow vectors shown indicate the first and second principal component eigenvectors. The first eigenvector pointing top-right lies on the direction of greatest variance as seen within the distribution whilst the second eigenvector lies orthogonal to the first indicating the direction of second greatest variance.

the dataset as possible. Subsequent observations can thus be transformed into eigenspace using the eigenvector matrix W, with lower order eigenvectors removed.

3.2.4 Linear Discriminant Analysis

LDA, one of the oldest classifiers still in use, is a supervised dimensionality reduction technique. Similar to PCA, LDA produces principal components or linear discriminants representing transformed axis vectors. However, whilst PCA, a non-supervised methodology, attempts to find eigenvectors which seek to maximise the variance within a distribution, the supervised LDA seeks to find linear discriminants which maximise the ratio of between-class distance $\tilde{\Sigma_B}$ to within-class class variance $\tilde{\Sigma_W}$

$$\arg_V \max L(V) = \frac{|\tilde{\Sigma_B}|}{|\tilde{\Sigma_W}|} = \frac{|V^T \Sigma_B V|}{|V^T \Sigma_W V|} \tag{3.7}$$

where V is the matrix of eigenvectors. As such, eq. (3.7) can be rearranged into a classic eigenvalue problem using

$$S_W^{-1} S_B V = \Lambda V \tag{3.8}$$

where V is the matrix of eigenvectors and Λ the diagonal matrix of eigenvalues which can thus be solved through eigen-decomposition to arrive at the linear discriminants.

Similar to PCA, dimensionality reduction can be performed by thresholding out the lowest ranked linear discriminants. Classification is also possible through a technique similar to Naive Bayes Gaussian where class probabilities along linear discriminants are calculated with new observations being classified based on the maximum *a posteriori* decision rule where the highest probability class is used as final classification decision.

3.2.5 Naive Bayes

naive Bayes (NB) is a family of probabilistic classifiers based on Bayes' theorem. Unlike other machine learning methods, NB requires no iterative parameter estimation, minimising computational complexity; whilst also having a linear scaling in parameter count versus feature count, minimising model complexity. The simplicity of such a model is however reliant on the assumption of strong independence between all features, a rare occurrence.

Let x_i be a feature where $x_i \in x = (x_1, \ldots, x_n)$. The naive Bayes' probability model is formulated as

$$p(c|x) = \frac{p(c)\prod_{i=1}^{n} p(x_i|c)}{\sum_k p(c)p(x|c)} \tag{3.9}$$

where the prior probability of class c is $p(c)$, while $p(x|c)$ is the probability of feature vector x given class c. The denominator, $p(x)$ is subsequently a scaling factor indicating the probability of feature vector x across all classes.

Traditional Naive Bayes Gaussian method of calculation is used on continuous valued features with the assumption of a normal distribution, whereas binary feature data can be measured using the Bernoulli naive Bayes model. The simplicity of Naive Bayes remains a primary advantage of such a classifier approach; whilst ironically, limiting overall effectiveness through such simplicity.

3.3 Deep Learning

DL based modelling approaches have an extensive history as the state-of-the-art in complex, big data driven ML applications such as computer vision, natural language processing and speech recognition[5]. Such popularity is achieved through the capability of automated representation learning—the ability to, itself, compose representations of raw data and consequentially model complex associations between features and outcome. The capability of automated representation learning minimises the need for hand-crafted feature engineering or significant data pre-processing, instead enabling the model to construct its own generalized feature representations from the provided raw data. As model depth increases, increasingly abstracted representations able to model non-linear and complex feature relationships. Accordingly, there is no surprise DL approaches are increasingly utilized within the similar big data domain of EHR based research[28].

The foundational modelling technique behind DL methodologies is the artificial neural network (NN) composed of fully-connected layers of artificial neuron cells or 'perceptrons'.

Before approaching the diverse domain of DL based architectures originating from the NN, we will review the fundamentals of the NN. With said foundational knowledge, we highlight the principles and current application of DL architectures followed by a closer look at the specialised domain of time-series based DL architectures.

3.3.1 Fundamentals: The Neural Network

NNs are a diverse and robust ML application able to model a variety of input-output mappings through the use of various network architectures. NNs have been argued to be able to equate to any optimal statistical classifier[73]. As such, NNs have shown great promise and continued use in various disciplines and domains[74, 75, 76, 77, 78, 79].

Nevertheless the capabilities of NNs come with the price of model complexity, with model parameters exponentially increasing with feature count and capacity. As a result, overly complex NNs suffer from long training times and issues with overfitting without a large enough dataset to match network capacity or the adoption of regularisation techniques[80]. Such high model complexity also suffers from the inability to validate trained models past that of empirical evidence from testing as opposed to theoretical validation. The black box nature of NNs limits the capability of understanding how a set of features and parameters are able to model a complex problem, only that it is able to.

NNs consist of sets of interconnected nodes arranged in layers called multilayer perceptron (MLP). Each MLP maps multiple input signals a_j^{l-1} through an activation function

$$a_k^l = \sigma(w_{jk}^l a_j^{l-1} + b_k^l) \tag{3.10}$$

to form a singular output a_k^l. Each input signal is modulated through a weighting w_{jk}^l and bias b_k^l before being aggregated into an activation function σ. Various σ functions exist with various properties and uses, the logistic sigmoid function being one of the most common.

Learning within a NN is based on the adjustment of weight and bias parameters in a feed-forward and back-propagation process. The forward pass consists of passing observations through a network consisting of a randomly initialised set of weights to generate an initial prediction. Through the use of a cost function, such as the mean squared error

$$C = \frac{1}{2n}\sum_{i}^{n}(y_i - \hat{y}_i)^2 \tag{3.11}$$

where n is the number of training observations, y the ground truth and $\hat{y}$ the model output, a loss can be formed on the distance error of prediction from ground truth. By minimising the

cost and subsequent model parameters, a NN can be pushed towards modelling the problem space and subsequent classification. Many algorithms exist which enable cost minimisation. An example of one being the stochastic gradient descent algorithm providing the capability to back-propagate changes in cost back through the NN updating weights and bias.

Stochastic gradient descent involves iteratively stepping down a hyperplane formed by the cost function and model parameters towards zero error or, more generally, a local minima which closely approximates the correct output. The direction of descent is determined through the solving of the pre-defined partial derivatives of the cost function. The model error eq. (3.12) and subsequent layer errors eq. (3.13) can be determined by:

$$\delta^L = \frac{\partial L}{\partial \hat{y}} \odot \hat{y} \tag{3.12}$$

$$\delta^l = (W^{l+1})^T \delta^{l+1} \odot \vec{a}^l \tag{3.13}$$

where $\frac{\partial L}{\partial \hat{y}}$ is the cost function partial derivative, $\vec{a}^l$ is the activation vector of layer l and W^{l+1} is the weight matrix of layer l. The direction of travel for layer parameters weight and bias can thus be calculated by

$$\frac{\partial L}{\partial b_k^l} = \delta_k^l \tag{3.14}$$

$$\frac{\partial L}{\partial w_{jk}^l} = a_k^{l-1} \delta_j^l \tag{3.15}$$

where b_k^l is the bias parameter for MLP k in layer l, a_k^{l-1} is the activation signal for MLP k in layer l and w_{jk}^l the weight parameter for MLP k in layer l to MLP j in layer $l-1$.

Finally parameters are updated by iteratively stepping towards a minima based on eq. (3.12) & eq. (3.13) through the following equations

$$w_{jk}^l \rightarrow w_{jk}'^l = w_{jk}^l - \frac{\eta}{n} \sum_i^n \frac{\partial L_i}{\partial w_{jk}^l} \tag{3.16}$$

$$b_k^l \rightarrow b_k'^l = b_k^l - \frac{\eta}{n} \sum_i^n \frac{\partial L_i}{\partial b_k^l} \tag{3.17}$$

where η is the learning rate indicating the length of stride for each iteration, n the number of observations, b_k^l the bias parameter for MLP k in layer l, w_{jk}^l the weight parameter for MLP k in layer l to MLP j in layer $l-1$ and b' and w' the future bias and weight parameter.

3.3.2 The Deep Neural Network and Beyond

The application of a traditional shallow NN itself is not considered DL; rather, DL represents a family of high capacity, multi-layer methodologies directing composition of feature repre-

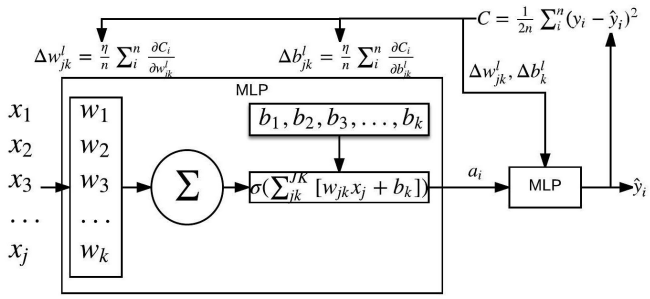

Figure 3.3: Diagram of a NN and multilayer perceptron with input, hidden and output layer. Input is passed through the network as a forward pass for error calculation before being backpropagated through the network to update weights and bias.

sentations towards modelling specific spatial domains. The simplest and most generalisable of DL methods being the deep neural network (DNN), simply consisting of multiple layers of fully-connected perceptrons enabling unrestrained composition of feature representations. Such generalisability is of great advantage, theoretically able to compose almost any optimal feature representation given the raw data. However, standard DNNs suffer from having such a large un-regularized optimization space containing many significant regions of local minima far-removed from the ideal global optimization minimum. Subsequently, effective model optimization purely through raw data and traditional training methodologies aimed correctly towards the global minima is a difficult prospect without leveraging *a-priori* knowledge into engineering limits or regularization into the optimization space to encourage correct optimization direction.

Within the state-of-the-art research field of engineering novel DL modelling methodologies, the application of *a-priori* knowledge to limit and regularize optimization spaces have spawned many DL architectures and techniques ideal for representing specialised spatial domains. Examples of such DL architectures include the convolutional neural network (CNN)[81], able to produce feature embeddings emphasising spatial location information of data points common within data domains utilizing euclidean space such as computer vision applications. Longitudinal information, involving data progression over time, is commonly modelled using recurrent type DL architectures such as the recurrent neural network (RNN)[82] and long short-term memory (LSTM)[83]; such models, able to 'remember' feature representations across time samples, enable past longitudinal information to be carried across a time-series to inform later model decisions. Irregular data domains involving non-euclidean space, such as transportation or communication networks, implement graph neural network (GNN)[84] type

architectures which generalize said spatial *a-priori* information into the graph domain via 'spatial' or 'spectral' transformation approaches.

Of interest, success within the aforementioned architectures are not strictly limited to specified domains or each applied exclusively within a DL model architecture. For instance, the CNN has shown considerable success within longitudinal domains[85] where the time-series component is considered an extra euclidean dimension whilst time-series based computer vision applications (such as video recordings) highlights the capability of a combined CNN and RNN architecture[86]. Non-standard applications of such architectures are also highlighted within this thesis in chapter 7 through the novel application of CNNs as a feature ranking mechanism.

Proceeding on, we highlight the foundational principles of several of the aforementioned DL architectures with relevance towards this thesis.

3.3.3 Autoencoders

Autoencoders are a form of deep NN with a distinctive architecture. Specifically, autoencoders have a generally mirrored architecture with input and corresponding hidden layers (called the encoding layers), reflecting output and corresponding hidden layers (called the decoding layers). Training objective is thus to output observations as close to what was input into the model. In order to stop encoding and decoding layers being trained into an identity mapping, layer capacity can be reduced to less than that of the input layer or through the use of regularisation techniques. Consequently, models are forced into learning a smaller and deeper representation of the data contained within the encoding layers. The encoding layers can thus be used as a feature encoder within a larger ML application much like PCA or LDA feature encoding.

3.3.4 Convolutional Neural Networks

CNNs are able to perceive spatial relationships between neighbouring features, such as images, unlike regular ML models in which flattened spatial data features lose all association with neighbouring features[6]. CNNs include the addition of convolution layers and pooling layers.

Convolution layers consist of convolutional filters which sweep through an input, applying a convolution operation to generate a feature vector. These convolution filters support update through back-propagation allowing for a number of features to be jointly represented by a small set of parameters reducing the size of what would be a large number of parameters in a regular NN. Said filters also represent neighbouring relationships by only convolving features within

its receptive field. Due to sweeping convolutional filters generating potentially increasingly larger model parameters, pooling layers combine outputs of multiple clusters into a single output through various strategies. Some of which include max pooling or mean pooling, outputing the max or mean value of all inputs respectively.

Combinations of convolution and pooling layers can be stacked to generate a deep structure. A prediction component is attached to perform final prediction based upon the learned encoded features of the above convolution and pooling layers.

3.4 Time-Series based Deep Learning

As previously mentioned, a primary objective of this thesis is DL based modelling approaches on time-series EHRs. As such, particular emphasis is placed on time-series based DL architectures. Consequently, we proceed to present an in-depth view of time-series based DL and highlight the challenges and solutions towards exploiting the unique data relationships within longitudinal data.

Longitudinal or time-series data consists of sequential snapshots of continually evolving measurements across time. Consequently, time-series data pertains a natural temporal ordering or, in other words, a dependent relationship between observations across a time period. The unique property of a temporal relationship between individual samples is lost within non-recurrent DL architectures with the assumption of independence between samples with no natural ordering of observations.

Time-series data can be worked around within strictly non-temporal DL architectures through data aggregation using pooling functions across the entire temporal dimension such as mean, max or min; or convolutional sliding window functions producing partially aggregated independent samples such as moving averages; however, such methodologies result in unrestrained loss of potentially significant temporal information. Alternative solutions include concatenation of entire series of samples with a reliance on temporal relationships being discovered during model optimization and composition via feature representation; there is however, no guarantee of such a temporal feature representation being composed naturally.

Of interest, previous mention of CNNs as an alternative successful architecture through representing the temporal dimension as a spatial dimension produces significant results in previous literature[85]. The clear association between correlation within nearby temporal samples and the relationship between nearby spatial points assumed by the CNN affirms said effectiveness within time-series data. This being the result of the adaptability of CNNs in localised

filtering across any dimension; be it the traditional spatial dimension or in this case: temporal. However, there remains drawbacks such as the limited receptive field of CNN kernels akin to having limited memory of previous temporal samples unlike with recurrent architectures such as the LSTM; whilst the natural property of greater temporal correlation between recent samples as opposed to distant is lost within the CNN, with no such feature representation assumption without human intervention and the use of handcrafted convolution kernels.

Proceeding on, will be a discussion of the two major foundational architectures used within the novel works of this thesis: the RNN and the LSTM. With which, the longitudinal property of EHR data can be exploited for improved patient outcome prediction.

3.4.1 Recurrent Neural Networks

RNNs are an adaptation of the MLP allowing for an internal memory state to be retained between observations. Consequently, RNNs are applicable to tasks involving time-series based data by "memorising" states from a previous time-step for use in a future prediction. RNNs include the addition of a weighted time-delayed recurrent connection which feeds a MLP output back into the MLP as an additional input. Said time-delayed recurrent connection enables a weighted output state to be stored and later included as a feature in future time-steps. Due to issues of vanishing and exploding gradients within the original implementation of RNNs limiting the availability of long term state memory, the LSTM unit was proposed using separate input, forget, update, and output gates to form a single node.

3.4.2 Long Short-Term Memory

LSTMs are a further adaption of the RNN concept. The use of weighted recurrent time-shifted connections, called the cell state, C_t, allows for the maintaining of memory of previous timestep samples. Through which, previous timestep data embeddings can be considered within the modelling and prediction of the current timestep. Figure 3.4 provides a diagram of individual components within an LSTM cell which will be briefly touched upon.

The update procedure of the past cell state, C_{t-1}, is controlled by the update gate which is formed of both the forget and input gate, $f_t C_{t-1}$ and $i_t C'_t$ respectively, by:

$$C_t = f_t C_{t-1} + i_t C'_t \tag{3.18}$$

Removal of embedding components within the cell state is dictated by the forget gate, comprised of parameters and σ activation function, forming a filter function learned through the

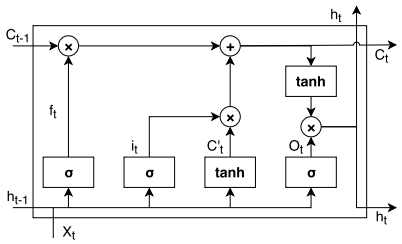

Figure 3.4: A diagram of a singular LSTM cell. As seen, the LSTM is comprised of multiple activation, weight pairs to form the four components intrinsic to the LSTM. Input data is passed into the LSTM to the input gate controlling data transformation into an embedded state for the update and forget gate to modify memory cell state. The output gate takes in said modified cell state in addition to the input data to produce the final activation output of the LSTM cell.

aforementioned LSTM input vector. Said bounded forget filter, $f_t \in \{0 \leq \mathbb{R} \leq 1\}$, is passed back to the update function eq. (3.18). The input gate is a combination of both input encoding, C'_t and selective filtering, i_t of the LSTM input to determine relevant encoding components to be incorporated to produce the new cell state. The activation output value, determined by the output gate, incorporates all previous components: the updated cell state, incoming input and activation from previous timestep to produce the final output of the LSTM, h_t.

$$h_t = O_t \cdot tanh(C_t) \tag{3.19}$$

$$O_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \tag{3.20}$$

Let w and b, be learned parameters unique to each LSTM component. Node input consists of x_t the input vector from the previous layer, concatenated with the activation output of said cell from the previous timestep, h_{t-1}. As shown, output is simply a learned filter function of the LSTM cell state, bounded by the tanh function. Accordingly, the number of learned parameters, w and b, increase significantly as compared to traditional NNs. Such increases however, are outweighed by proven significant improvements towards time-series based modelling applications.

3.5 Feature Selection

Within big data applications such as EHRs, the low prevalence of a small subset of relevant conditions within a large population cohort results in data that is high in both sparsity and dimensionality. Whilst DL architectures encourage and rely on large, high dimensional raw data to autonomously produce effective feature representations; the practical application of

such models within the domain of healthcare is infeasible. The manual entry of a complete, detailed, patient history, often unrelated to the current health issue, in order to produce a reliable model prediction is impractical. Additionally, the opaque "black box" nature of standard DL architectures produces model diagnoses without an underlying understanding or justification to a given model output; undesirable within the high-risk, high-repercussion field of healthcare. Consequently, applied patient diagnosis emphasises simplified decision flows using small subsets of significant biomarkers.

With said concept of simplification in mind, we look towards emphasising embedded feature selection applications within standard DL architectures. Through which, we seek to leverage automated feature representation composition to highlight significant predictive features of a specified patient outcome; discovery of biomarkers. The application of DL based feature selection enables complex non-linear associations to be discovered not previously available within traditional statistical selection methodologies. As such, we highlight the current overall state of feature selection methodologies as a foundation for future novel feature selection works within this thesis.

Feature selection operates under the assumption of feature redundancy or irrelevance within a large dataset; enabling the removal of said features without significant detriment to overall data informativeness: dimensionality reduction.

Of note, feature selection maintains feature independence between the mapping of old to new reduced datasets. As such, feature selection methodologies are distinct to the similar dimensionality reduction approach of feature extraction, producing transformed feature vectors dependent on the complete set of previous features. Consequently, the removal or selection of transformed features results in partial information loss within the original and thus a non-invertible function. Of importance, the subsequent mapping of importance metrics during analysis can not propagate back to the original feature set, or in our circumstance, distinct biomarkers; incompatible with the objective of biomarker discovery. Accordingly, focus will remain on solely feature selection approaches.

3.5.1 Wrapper & Filter Feature Selection

The field of feature selection can be separated into three distinct fields of filter, wrapper, and embedded applications dependent on the evaluation metric and sub-set selection algorithm used. Wrapper and filter type selection methodologies will be discussed below whilst embedded feature selection, a primary objective focus of this thesis will be discussed in detail

separately.

Wrapper

Wrapper methodologies, analogous to hyper-parameter optimization meta-algorithms within ML, use model predictive performance as the evaluation metric of choice for feature selection. A chosen model architecture is repeatedly trained on algorithmically chosen feature sub-sets of the overall data to produce model performance metrics. Feature sub-sets can be chosen via random bootstrap sampling, greedy forward selection or backward elimination of features, etc. dependent on the chosen filter methodology. Importance of said feature sub-set is thus inferred by the difference in model performance between sub-set trained model and a known baseline, either a hold-out validation set or against other sub-set trained model performance metrics. Wrapper methods, through direct optimization of model performance metrics, usually produce the best performing feature set for the applied ML methodology and modelling objective; however, results in significant computational complexity by the training of said ML method multiple times. Of note, an exhaustive search type feature sub-set selection via wrapper method would theoretically produce the ideal feature set; however, would be computationally infeasible (NP-hard on dimensionality) within a big-data domain using computationally expensive DL architectures.

Wrapper type feature selection methodologies remain in common use within recent literature for EHR data mining approaches[87, 88, 89]. As previously highlighted, computational complexity dictated analysis of only a small pre-mediated sub-set of potential predictors via wrapper methodologies within the aforementioned studies. Selection of prior sub-set of predictors in each study were, in fact, performed via less computationally complex filter type selection methodologies.

Filter

Unlike wrapper and embedded type feature selection methodologies, filter methods are applied without requirement for a ML predictive model; instead indicating feature importance via traditional statistical data informativeness metrics as a prior data pre-processing step. Features are ranked based on relevant statistical informativeness metrics against remaining features or experimental variable before selection based on a thresholding technique. Consequently, filter type methodologies are generally the least computationally complex as compared to wrapper or embedded methods requiring model optimization and input from performance metrics;

the most computationally intensive component of a predictive modelling pipeline. As such, simplistic filter methods sit ideally as fast selection applications within domains containing significantly large feature sets infeasible for wrapper type methods.

The concept of filter approaches to feature selection have a long history[90]. Filter approaches focus upon evaluating feature informativeness for feature ranking, of which there exists a large selection of varying metric properties. Common linear correlation approaches such as Pearson's R or Relief[91] enable the evaluation of linear correlation between feature and class label to estimate feature relevance towards future model prediction, named feature relevance approaches. Whilst feature redundancy approaches apply mutual information techniques such as Kullback-Leibler (KL) divergence to enable the assessment of feature-feature linear correlation. Application of both to ensure minimum-redundancy-maximum-relevance as proposed by Peng *et al.* [92] is a classical filter feature selection criterion. The aforementioned methodologies all fall under the category of univariate feature selection, the assumption of only singular-feature associations to class label as opposed to multi-feature associations. Multivariate feature selection remains uncommon within a purely filter feature selection approach, instead being of significant relevance within embedded feature selection applications[90, 93, 44].

3.5.2 Embedded Feature Selection

Embedded feature selection remains a dynamic, continuously evolving field of research within a large variety of big-data type domains including medical informatics[88]. Embedded type selection approaches incorporate both feature selection and model optimization as a unified singular objective within a ML algorithm. In doing so, embedded selection approaches take advantage of select ML algorithms' natural internal tendency for feature emphasis and selection during model optimization. The analysis of optimized model parameters enables propagation of importance of individual or sub-set of features in model predictive performance.

Embedded applications encompass a large domain of feature selection approaches spanning over a long period, being such a generic concept of unified optimization and feature selection. As such, within this section we lightly discuss a small subset of embedded feature techniques relevant towards the novel feature selection approaches proposed within this thesis. Accordingly, we introduce forward-backward methods, and sparsity term regularization methods.

3.5.2.1 Forward-Backward Methods

This subset of embedded feature selection methods behave in a manner similar to that of wrapper type approaches in that features are iteratively added or removed in a greedy manner,; however, as opposed to selection of features to optimize a generic ML algorithm's performance metric, forward-backward methods instead select weighted features in an attempt to approximate the minimization problem solution.

A variety of common embedded approaches follow said concept through iterative forward selection, backwards elimination, or a combination of both, defined as nested methods[94]. The classical, diverse, linear methods for regression exemplify forward-backwards embedded approaches to feature ranking. Forward stepwise linear regression approaches such as the traditional least squares[95] represent univariate feature selection whilst Gram-Schmidt orthogonalization[96] or the more modern least angle regression[97] represents multivariate regression approaches. Backwards elimination or shrinkage methods for regression include ridge and lasso approaches emphasising $L2$ and $L1$ coefficient normalization respectively[96]. Particular emphasis is placed on lasso, able to produce zero value coefficients via $L1$ normalization constraints, thus enabling continuous subset selection; consequently, lasso is both a backwards elimination and sparsity term regularization type approach. Sparsity term regularization approaches are detailed further in section 3.5.2.2.

Outside of linear regression approaches, decision tree type approaches such as CART[98], ID3[99], and C4.5[100] apply forward stepwise selection of features through recursively separating data based off an ideal feature at each node. Said ideal feature is dictated by importance measured by mutual information between features, i and output, y; where H represents entropy measurement:

$$MI(x^i, y) = H(y) - H(y|x^i). \tag{3.21}$$

Of interest, the RF prediction model (an extension of the decision tree approach) represents a non-forward-backward approach to feature ranking through the application of ensemble decision trees trained on bootstrap-aggregated (bagging) feature sub-sets with ranking metric derived by out-of-bag evaluation. Feature subsets are selected by random sample with replacement per ensemble decision tree as opposed to the greedy step-wise feature selection concept of forward-backwards methods.

Later approaches such as the popular recursive-feature-elimination SVM (RFE-SVM) proposed in 2003 by Guyon *et al.* [101] apply greedy backward selection embedded within the SVM classifier. RFE-SVM iteratively removes dimensions which decreases the separation hy-

perplane margins the least until σ_0 features remain where margin distance is dictated by $|w_i|$, the magnitude of model weight parameter w of feature i within the linear SVM case. Non-linear SVM feature selection is dictated by similar weight vector parameter α via:

$$W^2(a) = \sum \alpha_k \alpha_l y_k y_l k(x_k, x_l) \tag{3.22}$$

where the support vector margin is inversely proportional to $W^2(a)$ which equals $||w||^2$, thus the feature with smallest $W^2(a)$ change is selected for elimination.

3.5.2.2 Sparsity Term Regularization Methods

Within the case of linear models, feature selection can be approached by the regularization of model parameters to produce sparse feature representations; as opposed to greedy stepwise approaches, in which feature importance indicator metrics directly dictate the addition or removal of feature subsets. Said regularization is generally the addition of a sparsity penalization term to the model objective function.

Consider the traditional optimization approach of a linear model, $f(x_k) = wx_k + b$ attempting classification of a binary target, $y_k = \{-1, 1\}$. Optimization is approached as the minimization of the objective function:

$$\arg\min_{w,b} \frac{1}{N} \sum_{k=1}^{N} L(wx_k + b, y_k) + \lambda \Omega(w), \tag{3.23}$$

composed of $L(f(x_k), y_k)$ as the loss of the linear model prediction on training sample x_k compared to ground truth y_k; defined by various functions such as hinge, logistic or sum-of-squares. Being a quadratic programming problem, optimization of an ideal solution $f(x)$ based off prediction loss determined by empirical learning from a finite dataset is an under-determined problem space with a large region of possible solutions containing unclear local minima solutions. Regularization terms, $\Omega \in \mathbb{R}_+$, seek to restrict said solution space across a variety of favourable notions; common ideals such as smoothness or bounds of normalized vector space—otherwise, within a Bayesian viewpoint, imposing certain distributions based on prior knowledge. Said restriction is balanced by the λ coefficient, dictating emphasis between model prediction error and sparsity penalization.

In our case, sparsity regularization is imposed to restrict solution space to ideally minimize non-zero elements within weight parameter w. The zeroing of elements enables the effective elimination of features within the model function. Minimization of non-zero elements can be alternatively expressed as: $\arg\min_w ||w||_0$, minimizing the L_0 norm of w. Being non-convex,

optimization via quadratic programming of the l_0 norm is difficult[71]. Of interest, the solution to the L_0 regularized learning problem within a linear model is instead, equivalent to a wrapper based feature selection approach[102]. Instead, sparsity regularization or optimization of the L_0 norm is solved via approximation by the L_1 norm:

$$\Omega(w) = \sum |w_i|. \tag{3.24}$$

Element sparsity is encouraged based on the region of constraint produced through the L_1 norm. In reference to fig. 3.5, representing a theoretical parameter space, w, of two features, x_1 and x_2, under a regularized optimization problem; shown is the ideal unregularized model solution $\hat{w}$ and constraint regions, coloured blue, dictated by the L_1 and L_2 norm. The constraint regions for L_1 and L_2 are the diamond, $\sum |w_i| \leq \lambda$, and circle, $\sum ||w_i||_2 \leq \lambda$, respectively where λ represents the balance region between regularization and model loss, shown as red contours receding from ideal $\hat{w}$. An ideal regularized solution is subsequently the intersect point between contour and regularization constraint region. As seen, L_1 norm constraint regions produce protruding corners; if the ideal solution occures at said corners, parameters w_1 or w_2 are equal to zero thus removing the corresponding feature. Such corners enable greater likelihood of intersect between the two loss regions unlike that of L_2, having equal probability of intersect at any w_1 and w_2, being a disk. With parameter spaces of greater than two dimensions, the diamond becomes a rhomboid, providing more protrusions at locations $w_i = 0$ allowing for greater likelihood of parameter optimization solutions with zero weights.

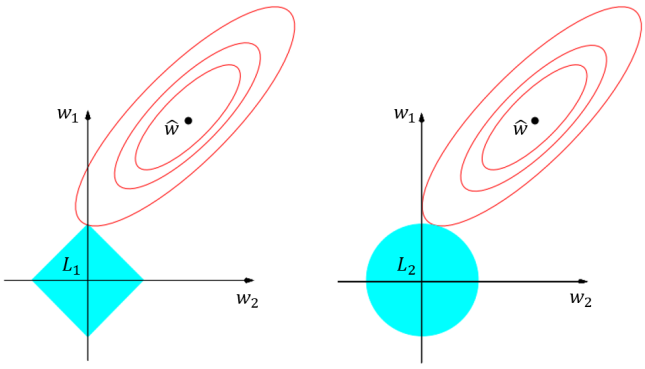

Figure 3.5: Estimation picture of a linear model parameter solution space containing two model weights, w_1 and w_2. Also shown is the ideal model solution for an unregularized model, $\hat{w}$. In addition the regularized constraint region $\Omega(w) \leq \lambda$ are shown in blue for the L_1 and L_2 norm, left and right respectively. As seen, the ideal balanced intersect between the two loss functions has greater likelihood of parameters equal to zero in the L_1 norm as opposed to the L_2; enabling sparsity normalisation within a linear model.[71]

Chapter 4

Applied EHR Modelling: An Overview of Clinical Objectives

Contents

4.1 Introduction

Electronic health record (EHR) technologies enable unrealised potential for large-scale, longitudinal analysis of a large selection of medical topics benefiting from retrospective, data-mining type health informatics applications. To briefly mention section 2.3, there exists a large selection of available opportunities to apply state-of-the-art machine learning (ML) based feature selection approaches to benefit individual health-based outcomes. Within this thesis, we will focus on two clinical aspects which remain a significant contributor to reduced patient outcomes: sepsis development within a intensive care unit (ICU) critical-care based setting; and dementia, a chronic and degenerative condition manifesting gradually over the course of years, prevalent within long-term care settings.

Within an informatics based viewpoint, sepsis and dementia present highly unique data characteristics within the over-arching field of EHRs. Sepsis—highly prevalent within the fast-paced field of ICU care—presents as a time-critical, comparatively short-duration, and high-frequency clinical objective requiring immediate diagnosis and aggressive treatment. Within the ICU, patient data such as vital signs are recorded at a frequency of potentially seconds to hours, with a more restricted set of medical events relevant in said ICU setting. Whereas, dementia exhibits as a highly chronic and gradually degenerative condition, manifesting over the course of years within the infrequent setting of primary care institutes such as general practice (GP), hospital outpatients, and the occasional hospital spell (for the treatment of, or for a co-morbidity as a result of, dementia). Consequently, patient data expands to a highly sparse and varied set of medical events with inconsistent frequency and long-duration across a large selection of health-care institutes. The application of such diversely characterised clinical fields serves to demonstrate and validate our novel ML contributions within this thesis to demonstrate model robustness and diversity.

We proceed to discuss in detail, highlighting individual motivations, challenges and current state-of-the-art within the distinct health informatics fields of sepsis and dementia prediction for the remainder of this chapter.

4.2 Dementia

Through continued modern advancement, life expectancy has steadily increased whilst fertility has decreased resulting in a continually ageing average population requiring continuously evolving oversight[45]. One such significant challenge of an ageing population is the diagnosis,

treatment and continual care of chronic and degenerative cognitive decline, the manifestations of dementia. The impact on patient care givers, families, and society is substantial[103, 104], whilst the significant number of cases globally is only predicted to increase steadily[105].

Being a chronic and degenerative condition, dementia has a high prevalence within the elderly affecting 47.5 million people globally[106]. The most common cause of dementia being Alzheimer's Disease (AD) making up 60-80% of cases[107]. Within the UK, dementia affects 850,000 people with a forecast rate of prevalence of 1 million by 2025 and 2 million by 2051[103]. The resulting cost of dementia in the UK totals £26.3 billion with two thirds being paid for by dementia patients and families in private social care as of 2018[103].

Initial diagnosis of dementia is highly reliant on relatives or self reports[108], confounded also by prevalent co-morbidities in the elderly[109]. Current certain diagnosis generally requires a battery of clinical tests, such as cognitive assessments, patient history questionnaires and neuroimaging; whilst other potential causes must be ruled out for a more conclusive diagnosis[110]. A truly definitive diagnosis is only possible through a post-mortem autopsy[111]. Consequential inconsistent or delayed diagnosis occurs late into dementia development[112] resulting in reduced effective care and potential patient outcomes[105].

EHRs, encompassing extremely long-term patient records serves as an ideal platform for our particular focus on analysing long-term biomarkers for chronic and degenerative dementia. Data driven ML techniques have the capabilities of modelling such complex associations, as proven within other fields[6, 113]. Making the best use of these big health related data, ML techniques provide a way of delivering high quality personalised healthcare services in real time. Within current literature, ML has demonstrated promising applications to neuroimaging analysis. Orru *et al.* [114] surveyed the application of support vector machines (SVMs), preceding 2011, in identifying imaging biomarkers of neurological and psychiatric diseases. Mosconi *et al.* [115] reviewed the existing scientific literatures involving the early detection of AD using neuroimaging. Mosconi *et al.* focused on the effectiveness of neuroimaging detection, possible risk factors and the progression from healthy to general cognitive impairment. Recent major reviews, such as by Ching *et al.* [116] provides overviews of ML within biology and medicine.

Following on, a brief overview of the unique combination of challenges within dementia care is presented, followed by an in depth explanation and evaluation of common ML methodologies used within dementia informatics. We then summarise relevant literature and bring forward potential unexplored avenues of research.

4.2.1 Challenges

Dementia presents a unique assortment of varying challenges in regards to diagnosis originating from dementia being a chronic and degenerative set of conditions. The detection and diagnosis of the initial stages of dementia or mild cognitive impairment (MCI) continues to be problematic[108] with a reliance on self reporting or reports by relatives. Compounded with symptoms being obscured by the regular effects of natural ageing, potential cases of dementia are generally reported 2 to 3 years after onset[110]. The delayed diagnosis results in continued unchecked decline which reduces the effectiveness of any care given upon discovery and diagnosis[117, 118, 119].

The current diagnostic model begins with several screening procedures used to identify potential dementia patients for further evaluation leading to a definitive diagnosis, the most common being the Mini Mental State Exam (MMSE) and the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) cognitive assessments[73]. However, current cognitive assessments remain problematic with the MMSE being called into question as an effective screening measure, Tombaugh *et al.* [120] provides a thorough review of literature evaluating screening effectiveness with a sensitivity varying from 21-100% and specificity of 46-100%.

Such a combination of factors results in a slow and ineffective system of diagnosis with an estimated two thirds of dementia cases remaining undetected[121, 122].

As mentioned previously, diagnosis generally involves the use of neuroimaging after screening to further assess the potential for dementia. The most common method being magnetic resonance imaging (MRI) and Positron-emission Tomography (PET). Current evaluation of scan results requires the use of an expert radiologist or anatomist in order to correctly identify and perform manual measurements. Such manual tasks often result in excessive time consumption, variability between different medical professionals and are limited to only certain brain regions[123]. With current manual MRI evaluation[124] the resulting separation between normal elderly and probable dementia achieves accuracies ranging from 58-100%. Kloppel *et al.* indicates a significant difference in diagnostic ability between general radiologists and neuroradiologists in evaluation of clinical trial[125] which may clarify the apparent variation in accuracy. Kloppel also uses the results as an indication of the need for specialisation. All of which, speaks for the need of more straightforward and effective utilization of neuroimaging data.

Several shortcomings exist with the use of neuroimaging applications; the foremost challenge being cost[126, 111]. Cheaper and more readily available alternative diagnostic proce-

dures for dementia may serve to address the problematic issue of scarcity and cost of specialist neuroimaging services.

Interestingly, social stigma also factors into the challenge of needing a supplementary diagnostic procedure for diagnosis. Boustani *et al.* [107, 127] reports high refusal rates of 51% of potential Dementia screen patients for further diagnostic assessment after a positive screening result. Boustani's findings also suggested that patients believed dementia to be a devastating condition with no available treatment and would lead to issues such as depression, anxiety, social stigma, insurance coverage and loss of independence. Changes to diagnostic procedure which allows for immediate effective diagnosis may alleviate the issues of patient refusal.

The prognosis of dementia is signified with continued degradation of mental ability, increased risk of co-morbidity[109] and increased risk of institutionalisation resulting in a significantly higher risk of death than non-demented patients[128]. The reduced health of dementia patients also produces a corresponding significant increase in healthcare costs[109]. The risks of acute conditions and events resulting in hospitalisation are also affected with an onset of diagnosis. Hospitalisation by falls remains a significant issue for the elderly[129]. The fall rate of nursing home residents with dementia nearly doubles compared to residents without in a study by Van Doorn *et al.* [130]. Further research is required in effective prognosis and care post-diagnosis to alleviate the significant increase in co-morbidity risk posed by dementia.

4.2.2 Current State-of-the-Art

There exists a large selection of literature regarding ML based detection of dementia. EHR based dementia detection presents potential for efficient hands-off automated detection across a large majority of the population based off critical indicators or medical events indicative of developing dementia[131, 132].

Within the reviewed literature, a consistent set of ML technologies is applied within the diverse fields of dementia diagnosis. As shown in Table 4.1, the distribution of underlying ML methodologies used indicates SVM as the popular methodology in use overall. With such diverse data domains available for the overall goal of dementia detection, there exists multiple avenues of analysis to arrive at an effective diagnosis. There further exists commonality in ML approaches applied in such distinctly unique data domains, highlighting the adaptability possible with even traditional, basic ML techniques.

The field of ML in dementia diagnosis has been very active over recent years with applications making use of a variety of patient data and methodologies for diagnosis with the overall

Table 4.1: Underlying ML methodologies used within the reviewed literature

ML model	Total Count	Literature
		Cognitive Assessment
SVM	3	[73, 133, 134]
NB	3	[73, 135, 134]
LR	3	[135, 136]
DT	2	[134, 135]
NN	2	[73, 134]
LDA	1	[136]
RF	1	[73]
		Neuroimaging
SVM	14	[12, 114, 125, 142, 143, 144, 145, 146, 147, 137, 138, 139, 140, 141]
LR	5	[12, 123, 142, 148, 149]
LDA	4	[12, 123, 140, 150]
PCA	2	[138, 148]
NB	1	[12]
DT	1	[12]
NN	1	[12]
RF	1	[142]
CNN	1	[10]
DNN	1	[11]
		Speech Assessment
NN	2	[110, 151]
LR	1	[151]
DT	1	[151]
NB	1	[152]
PCA	1	[153]

CNN: Convolutional Neural Network, DNN: Deep Neural Network, DT: Decision Tree, LDA: Linear Discriminant Analysis, LR: Logistic Regression, NB: Naive Bayes, NN: Neural Network, PCA: Principal Component Analysis, RF: Random Forest, SVM: Support Vector Machine

goal of discovering novel biomarkers for diagnosis or to improve upon diagnostic ability. Various papers have proposed the use of more novel patient data such as interview transcripts or EHR in an attempt to move away from the expensive use of neuroimaging[126, 111]. In regards to methodology, SVMs were the most popular ML model used within reviewed literature. Various other methodologies such as random forest (RF), linear discriminant analysis (LDA), Bayesian network (BN), principal component analysis (PCA) and occasionally neural network (NN) are used in support of or as diagnostic methodologies.

4.2.3 Opportunities & Future Research

Whilst the general field of big data analytics continues to mature, the current state of ML in dementia diagnosis remains behind current state of the art methodologies. Nonetheless many studies have proposed applications able to deliver promising dementia biomarkers or propose diagnostic procedures in collaboration with ML methods able to outperform current procedure.

Several avenues of research still remain relatively unexplored in addition to advances in fields of ML opening previously unavailable avenues.

4.2.3.1 Health Data Linkage

Reviews of current literature identify limitations of small validation case amounts, improved validation measures, and the need for diverse EHRs separate from the prevalent ADNI database[131, 154].

A major prerequisite for any big data based complex modelling and applications is data availability. With the majority of research currently relying on small patient groups with observations in the hundreds to occasional thousands, various organisations have devoted immense effort into the creation of large-scale datasets appropriate for research. The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset for MRI scans and patient demographics was mentioned previously along with the CERAD database for neuropsychological assessments. Several other databases exist, as shown in Table 4.2, for generic EHRs which provide extremely large, full featured datasets of patient history. Such datasets open a new avenue into dementia prognosis based on the chronic and degenerative nature of dementia providing continual data on individuals. Coupled with time-series modelling, EHRs enable the exploration of dementia prognosis.

Table 4.2: Several large scale EHR and data linkage databases.

Name	Region
SAIL[29]	Wales, UK
SHIP[155]	Scotland, UK
Data Linkage Western Australia[156]	Western AUS
ICES[157]	Ontario, CAN
MCHP[158]	Manitoba, CAN

EHRs however, provide multiple challenges which limit potential applications. The predominant challenge being the wide-ranging and non-specific patient information recorded in such datasets. The resulting patient data presented are generally sparse and highly dimensional, compounded by a lack of prior knowledge in what constitutes as relevant data utilised in specific domains such as dementia diagnosis. The use of sparse, high dimensional EHR data within health informatics presents two major challenges: human interpretability, requiring the employment of sparse optimised feature selection, dimensionality reduction or representation learning for effective biomarker and risk factor identification; and adequate data coverage pro-

ducing meaningless artefacts and bias termed *sparse data bias*, potential solutions of which exist[159].

In regards to EHR encoding, various avenues of research exist which address this challenge: representation learning technologies remain a constantly evolving field[160] with which to adapt into the field of EHR health informatics. Alternative methodologies from other ML fields allow for potential adaptation into EHR encoding such as word representation approaches within natural language processing (NLP), of which methodologies such as word2vec by Mikolov *et al.* remains highly popular[161]. With no single *de facto* methodology for EHR encoding, there remains great potential in the proposal of novel tailor-made encoding methodologies for EHRs.

Finally, relatively little research has focused on evaluation or diagnosis across simultaneously multiple data types. With health data linkage continuing to provide the possibility for full, structured and detailed records for individual care, the use of detailed assessments such as MRI, Electroencephalogram (EEG), and cognitive assessments can be coupled with long term patient histories from EHRs allowing for the creation of fully fledged and thorough diagnostic support systems. While such work has been attempted using statistical methods[107], and through ML methods[111], little research has continued within such research avenue.

4.2.3.2 Prognosis

Within reviewed literature, the classification of MCI versus dementia patients remains a continually challenging observation[162] with reported evaluation accuracies indicating a consistent significant drop in comparison to control versus MCI or full dementia classification[163]. The use of neuroimaging, cognitive assessment and discourse analysis have been unable to classify continued degeneration effectively whilst other approaches such as EHR remain unexplored[43].

In retrospect, little research has also gone into actual prediction of MCI and dementia conversion based on historical patient history. Such research applications would provide great potential into identification of risk factors and biomarkers indicating rates of cognitive decline. Several clinical studies have attempted equating cognitive decline to cognitive assessment scores statistically[164] however, the use of modern ML techniques may provide novel indications.

Continued cognitive decline provides a sequential time-line of discrete events indicating the gradual worsening of dementia symptoms, such time-series based data serves as a per-

fect example for modelling on time-series based methodologies. As mentioned in section 3.3, recurrent neural networks (RNNs) allow for short term memory of past events ideal for applications modelling dementia conversion. Such applications can potentially provide improved predictions of cognitive decline allowing for personalized tailored medical care or identify future at-risk individuals for close monitoring. However, deep learning (DL) technologies such as RNN remain a concern in human interpretability and validation. Consequently, following on from the example of at-risk identification, indications to the reasoning behind predicting an individual as at-risk remain unknown. Such issues, remain an ongoing research challenge.

The degenerative nature of dementia results in an increase in comorbidities[109], institutionalisation[128] and fall rate[130]. Several studies have proven the statistical significance of dementia as a risk factor to hospitalisation[165, 166, 109, 167]. There remains untapped potential in prediction of institutionalisation risk and hospitalisation outcomes for dementia patients using ML applications.

4.3 Sepsis

Despite significant modern advances in antibiotics and acute care management, the development of severe sepsis produces significant negative patient outcomes[168]. Sepsis arises from an overly extreme immune response to infection, causing significant injury to tissues and organs. With both a high prevalence and significant mortality rate[169], severe sepsis remains the primary cause of death from infections[170] resulting in significant concerns for practitioners within an ICU setting.

Current diagnostic procedure and treatment strategy places extreme importance in time-critical early detection and treatment of sepsis symptoms[171]. The current UK treatment strategy, the *Sepsis Six*, places emphasis on early intervention with a substantial treatment plan of IV antibiotics and fluids followed by intense continual monitoring within the first hour of sepsis suspicion before follow up confirmation of diagnosis by blood work[171, 172].

The pathophysiology of sepsis is still an uncertain prospect with established diagnostic procedure undergoing significant change over recent years. Historical definitions of general sepsis in the 1980s[173] evolved to the distinction of severe sepsis and septic shock and the systematic inflammatory response syndrome (SIRS) definition in 1991[174], later renamed the sequential organ failure assessment (SOFA) criteria. Current modern-day established criteria involves a recently developed simplified system of diagnosis called the quick SOFA (qSOFA) system, in 2016[170]. The qSOFA system greatly simplifies the SOFA, points based multi-

categorical scoring of vital signs from six individual biological systems, to just a simple 2 out of 3 positive indications of low blood pressure, high respiratory rate or an altered state of consciousness. QSOFA provides both a quick and simple clinical assessment procedure whilst maintaining effective discrimination of sepsis with respectable baseline AUROC scores of 0.72[170].

Within critical ICU based care, continuous monitoring of patients produces and records significant quantities of high-frequency patient vital data. Such EHRs serve as another ideal platform for analysis and automated detection of time-critical medical events such as the aforementioned development of sepsis. Current procedures involving manual monitoring through simplistic rules based diagnostic criteria serves to benefit from the application of automated detection involving ML based applications.

Following on, a brief overview of the unique challenges faced within sepsis diagnosis and treatment is presented, followed by a discussion of the current state-of-the-art in sepsis detection within the field of health informatics. Finally, a brief exploration of future avenues of potential research in addition to identifying several available unexplored opportunities to further enhance sepsis based patient outcomes.

4.3.1 Challenges

Specifically within UK statistics, prognosis of a septic patient indicates a 35% mortality rate during ICU stay[168], 47% mortality rate during hospital spell[168] and a 63% rate of hospital readmission within the 1st year[175]. Such a severe prognosis is additionally met with a high prevalence rate of 27.1% of adults meeting severe sepsis criteria within the 24 hours of ICU admission[168]. Such statistics provide a snapshot into the significant severity of severe septic development within a patient.

Emphasis on the need for time-critical intervention upon suspicion of sepsis is apparent; with a 5-8% increase in mortality per hour for sepsis and septic shock respectively when left untreated[171, 176]. Consequently, advanced early detection of sepsis development is necessary to ensure minimal patient mortality and subsequent improved medical outcome.

As mentioned previously, current approaches for septic patient identification revolve around simplistic clinical rules based upon detection of sequential organ failure such as the older SIRS system to the more modern qSOFA system. Such applications, able to produce effective baseline AUROC scores of 0.72[170], raise concerns regarding poor sensitivity within the qSOFA diagnosis system[177]. Inversely, the SIRS system instead suffers from poor preci-

sion[178]. The resulting delays in sepsis detection results in inappropriate antibiotic use[178].

Sepsis, being a condition predicated by infection whilst all infections not consistently leading into sepsis, results in significant challenges in clear definitions and indications of sepsis. Diagnostic criteria mentioned previously further focuses on detection of organ dysfunction as opposed to direct diagnosis of sepsis. As such, applications such as qSOFA and SIRS detect based on organ dysfunction predicated by sepsis predicated by infection, resulting in significant difficulties for absolute definitions of event start. Such complications manifest within a health informatics standpoint as highly varied case study approaches to sepsis definitions and indications within the literature. Case study data heterogeneity remains a significant challenge in effective literature comparisons of model capability. For instance, a large collection of studies follow the Sepsis-3 clinical criteria[170] as either:

1. Recorded two point deterioration in qSOFA score.

2. Indications of clinical suspicion as blood culture testing and 72 consecutive hours of IV antibiotic administration within a certain time period as follows:

    a) Prescribing of IV antibiotics followed by blood culture testing within 24 hours.

    b) Blood culture testing followed by prescribing of IV antibiotics within 72 hours.

or the earliest indication if both are present. There exists however, alternative clinical criteria applied by various literature based on the then established medical understanding of sepsis.

Compounding said lack of case study heterogeneity is variations in definitions of sepsis indication time. With current trends in literature aiming to outperform current diagnostic criteria, there also exists the trend of predictions at an earlier time-frame. As such, indication time objectives generally shift by upwards of 6 to 24 hours early within various studies. Modelling difficulty between 6 and 24 hours is significant with reported predictive evaluation metrics varying greatly[179, 13, 15]. With model performance reducing significantly with earlier detection time objectives (4 hour difference) by as much as 14% (AUROC)[179], consistent comparisons between studies is difficult.

Dataset selection presents challenges of case-control onset time matching to ensure population heterogeneity. As with any health informatics based study, case-control population matching can be applied via similarities in characteristics such as gender and age. Of greater difficulty within longitudinal studies is defining cut-off times for patient timelines. Positive populations can have cut-off times pertaining to just after sepsis onset, however there exists no

clear distinction within control populations. Inclusion of a complete control patient timeline results in generally clinically stable indications, as opposed to the underlying clinical indications present in sepsis positive patients, resulting in population bias. Such aspects must be carefully controlled with clear definitions of control population cut-off, an aspect lacking within the reviewed literature.

4.3.2 Current State-of-the-art

A review of current literature on ML based sepsis prediction applications highlights multiple recent directly related studies over a period of 2016-2019. Focus is placed on ICU based settings, presumably due to the greater ubiquity of high-frequency, detailed recordings of patient vital measurements in such an environment. Some exceptions exist, including Masino *et al.* [180] focusing on neonatal ICU units (<1 year of age) and Le *et al.* [181] using pediatric patient records (2-17 years of age). Recent literature has indicated the untapped potential of ML based screening applications. Various comparative studies show novel ML applications surpassing traditional screening approaches such as qSOFA, SIRS, etc. in performance[14, 181, 13]. With adoption of such ML algorithms within controlled real-life ICU settings resulting in significant improved patient outcomes[182].

Several studies apply direct comparisons of novel applications against the established SOFA system[13, 14, 15] or the similar SIRS criteria[13, 181, 183, 14, 184, 15]. All the aforementioned studies present significant performance improvements over SOFA or SIRS, highlighting the potential for augmenting sepsis detection through ML based monitoring applications. Significant studies include McCoy *et al.* [9], who presents an implementation of such a ML sepsis prediction application within a real-world hospital environment in replacement of the SIRS criteria. McCoy *et al.* reports a significant decrease in mortality rate by 60.24% in addition to improvements in both patient length of stay and readmission rates post-implementation.

Within the literature, ML methodologies span a large collection of traditional ML models which include: Faisal *et al.* [185] using classical logistic regression models on transformed features and Horng *et al.* [186] using SVM on encoded representations of text based records. Gradient boosted decision trees with various novel feature transformations have also seen use by Delahanty *et al.* [13] and Le *et al.* [181]. Van Wyk *et al.* [187] presents a comparative study of various ML methodologies resulting in RF outperforming other methodologies including: SVM, logistic regression, NN and RNN. The proprietary InSight all-in-one decision support

platform was highly popular with 5 studies comparing or directly applying InSight into the decision pipeline. One paper, by Kam *et al.* [184], involves the use of DL based methodologies.

4.4 Conclusion

EHRs present great potential for large-scale and long-term analysis of a broad variety of medical applications in aid of improved patient outcomes. As highlighted, the positive impact of EHR analysis extends to highly diverse and unique non-trivial applications as evidenced within this thesis.

The modelling of extended long-term patient outcomes of a continually degenerative condition such as dementia presents challenging problems such as highly disparate, inconsistent and incomplete health records. Health outcome objectives such as condition prognosis, risk-factor identification and institutionalisation risk analysis enables greater focus on preventative measures and care as opposed to reactionary mitigation of associated negative events within a degenerative and presently incurable condition. Current research and state-of-the-art in ML based analysis highlights a picture of diverse modelling approaches across a large variety of EHR types.

On the opposing side of medical application, sepsis within a clinical environment produces unique challenges of time-critical early diagnosis of a highly prevalent and severe condition. Confounded by common multi-morbidities which cause the initial hospitalisation event in addition to non-simplistic definitive diagnosis procedures, sepsis prediction is a challenging application domain. Continuous monitoring and recording of historical septic events raises the possibility of automated monitoring using improved ML based applications able to outperform that of current simplistic rules based clinical screening procedures. Current state-of-the-art focuses upon ICU based EHRs applied upon a small selection of possible recorded features able to out-perform current clinical approaches such as qSOFA or SIRS.

The contrast of both medical applications provides a diverse domain of potential approaches and applications in which to advance the state-of-the-art. There exists great potential for significant improvements in individualised patient care and outcome through the leveraging of ML based modelling techniques in a relatively young domain of EHR analysis. Following on, this thesis presents the significant developments and contributions towards the state-of-the-art in ML based longitudinal EHR analysis in dementia patient outcome modelling and advanced early detection of sepsis across three major chapters.

Chapter 5

Dementia Hospitalisation: Risk Factor Identification Using Entropy Cascades

Contents

5.1 Introduction

As discussed within chapter 2, with such a vast domain encompassed by the medical and social services potentially experienced by a patient, big data of such nature will invariably suffer from the *curse of dimensionality*, resulting in data domains consisting of upwards of thousands of dimensions. Consequent data sparsity follows behind as population size is vastly outpaced by the required sample size needed to maintain statistical significance for the size of feature space. Healthcare data poses a significant challenge for the traditional statistical approaches generally applied within health informatics [43]. The use of such data within general predictive machine learning approaches poses additional challenges on interpretability and application on a human level. Without a reduction of feature size to a manageable size, the practicality of such approaches will remain outside of medical application, and firmly within the confines of academic interest.

To address such challenges, this chapter introduces a novel initial application of an embedded feature selection and prediction methodology for hospital admission of individuals with dementia on a sparse, high-dimensional dataset of medical events. By performing feature selection in parallel with classification training, selection of features can be focused on identifying effective discriminative features relevant primarily to the required task at hand. Being generic electronic health records of patient history without any direct relationship to dementia analysis or diagnosis, the reduction of the hundreds of thousands of potentially unrelated medical events to only a handful minimizes the number of redundant variables in need of further clinical or statistical study in identifying potential risk factors. The collection of electronic health records via the Secure Anonymised Information Linkage (SAIL) data-bank [56] allows for the linkage of anonymized patient records across the various healthcare providers such as general practice (GP), in/out-patient hospital records, population deprivation, etc. This provides the potential of novel research applications involving the entirety of a patient time-line from birth to death.

Various studies have gone on to explore common causes of hospitalization within the population of dementia sufferers with a focus on clinical study and survey data with limited population scope as discussed in section 4.2. Kalisch *et al.* [188] identified, through a retrospective cohort study, a significantly increased risk of hospitalization for demented individuals when taking two or more anticholinergic medications with an adjusted incident rate ratio of 2.58. Chan *et al.* [189] follows a similar line of investigation indicating that 53.4% of cases of hospitalization of the elderly due to adverse drug events were preventable due to non-compliance

or omission of indicated treatments. Phelan *et al.* [165] identified causes of hospitalization such as bacterial pneumonia, congestive heart failure, dehydration, duodenal ulcer and urinary tract infection as being significantly higher among those with dementia. Naalwala *et al.* [190] provides similar conclusions while also including causes such as bronchopneumonia. Bynum *et al.* [109] provides a more extensive list of hospitalization causes whilst also identifying the number of comorbidities as a consistent association with the odds of hospitalization. Toot *et al.* [166] establishes factors such as behavioral problems including agitation and wandering, as well as changes in daily living routine, to have an increased risk of hospitalization for people with dementia.

While the studies mentioned have provided informative results, the resulting causes of hospitalization all refer to a root cause in hindsight of the actual hospitalization event. Little research has been performed on identifying influential risk factors and clinical events from previous health records in an attempt to predict patient hospitalization. Related fields of research such as dementia diagnosis decision support systems have seen comparatively greater interest in the use of big data machine learning (ML) approaches. The resulting methodologies created from such fields of study provide great opportunity for adaptation into data mining and risk factor analysis.

Within the overall context of this thesis, the proposed methodology serves as an initial adaption of multiple, established components with modification to produce a novel embedded feature selection approach applied within an unexplored domain; namely, electronic health records (EHRs) for hospitalisation risk modelling.

5.2 Methodology

The proposed method, entropy cascading neural networks (ECNN), consists of a four-stage pipeline: initial training using entropy weight regularization, snapshot ensemble training & aggregation, feature importance grouping & ranking, and backward-stepwise feature selection & validation for risk factor analysis. The proceeding section presents initial data preprocessing following on with individual pipeline components, examined in detail.

5.2.1 Data Preprocessing

ECNN emphasizes the use of patient records consisting of GP read codes over a time period of multiple years. More detail of the experimental dataset is presented in section 5.3. All unique

read codes were one-hot encoded as individual features with each patient sample indicating total occurrence of read code over the relevant time-period (see section 5.3).

Data preparation involved normalization on a per feature level with values linearly scaled to a maximum and minimum range of $[0, 1]$. With values being total occurrence within a set time-frame, missing data is thus a zero count within a certain feature, time-frame data value. Class labels were set as a binary label indicated by $\{0, 1\}$ as positive and negative instance respectively across the entire patient timeline. Said binary label indicates the presence of any hospitalisation event occurrence after initial diagnosis of dementia.

5.2.2 Entropy Weight Regularization

As mentioned previously, dimensionality and sparsity are the main challenges of data analytics using electronic health records. With data dimensionality potentially numbering in the hundreds of thousands and individual observations having perhaps tens of values, the leveraging of such data in producing an effective predictive model whilst maintaining comprehensibility is a hard prospect.

Traditional dimensionality reduction pipelines such as principal component analysis (PCA), relying on orthogonal transformations of the dataset, suffers on a comprehensibility standpoint. After said orthogonal transformation into the new embedding space, with axes not necessarily parallel to the original feature space axes and based off orthogonal vectors of most variance, each of the resulting orthogonal dimensions or principal components become fully dependent on every original feature. After the removal of low-variance principal components, the traditional methodology for PCA dimensionality reduction, a transformation back into original feature-space would result in information loss across multiple features due to the aforementioned dependence. Consequently, the selection of a single principal component of high-importance would transform into a vector spanning across the entire feature space. Subsequent selection or ranking of individual read codes for clinical significance would thus become highly impractical. A final application involving the use of such dimensionality reduction methodologies will still require the evaluation of every medical event within a patient time-line. Another major disadvantage of such methods is the apparent disconnect between dimensionality reduction and prediction. PCA bases dimensionality reduction on the variance of a dataset and as such performs reduction without any feedback as to its effectiveness.

The method proposed below seeks to solve both issues. By performing feature selection during the training of the predictive model, feedback on the performance of the predictive

model based upon the reduced features can be fed back into selecting features relevant to the trained task at hand. In addition, reduction will be performed directly on feature dimensions and as such, allows for the direct removal of redundant events within a patient time-line.

This paper proposes a novel adaption of the entropy regularization technique, originally proposed by Zhou *et al.* for support vector machine (SVM) models eq. (5.1), towards the neural network (NN) architecture. The measure of information entropy defines the potential information content of a data source or the unpredictability of a certain state occurring. As such, within a probability mass function, $P(x)$, of a binary variable, x, the information entropy of said variable will approach zero where the probability mass function approaches near certainty of one or the other action. The information entropy is highest at the midpoint, $P(x) = 0.5$, where the probability of either action is exactly equal. Consequently, this property of information entropy can be leveraged into enforcing weight sparsity within our methodology.

By incorporating entropy regularization based on the bounded weights of the first layer of the NN within the cost function, weight updates will seek to minimize entropy, thus driving said first layer weights towards $\{0,1\}$. The original cost function seeks to push weights in either direction towards improving predictive accuracy. With a linear activation function, weights approaching zero will filter out activation signals whilst weights approaching one will remain unaltered. Entropy regularization will emphasize the need to push weights towards boundary extremes. The combination of the aforementioned functions will result in activation signals of importance being driven towards one whilst redundant signals in the scope of predictive performance will be pushed towards zero and thus filtered out. The resulting weight matrix will be of a sparse form consisting of only activation signals which contribute to the model prediction. The entropy cost function is thus:

$$L(w) = -\lambda \sum_{jk}^{JK} w_{jk} \log(w_{jk}) \tag{5.1}$$

where w_{jk} is the weight representing the connected edge between the k-th multilayer perceptron (MLP) in layer l and the j-th MLP in layer $l-1$. The hyper-parameter, λ is a regularization coefficient to fine-tune the balance between predictive performance and weight sparsity. Consequently, weights close to zero will map to $\theta = 0$ while highly positive weights will map towards $\theta = 1$. The resulting sparse weight matrix of the first layer will act as a filter, removing inconsequential connections between MLPs within the first and second layer. By evaluating this matrix, the resulting input features can be categorized into three types shown in ascending order of importance:

Disconnected Features whose weighted connections have been driven close to zero are completely excluded from the remaining model and as such, are non-meaningful features for classification.

Partially Connected Features where only some weighted connections have been driven close to zero. Consequently, these features exhibit element-wise sparsity and as such remain partially used.

Fully Connected Features whose weights exhibit non-sparsity indicates a favorable feature which remains in use for the remainder of the model.

Redundant features are thus removed by selecting favourable features whose associated weights are fully or partially connected with large magnitude. Through associating feature selection based upon parameters within the predictive model during training, feature selection can be tailored towards selecting features which favor heavily into the overall predictive performance.

5.2.3 Snapshot Ensembles

The training procedure used involved the use of a modified snapshot ensemble training procedure proposed by Huang *et al.* [191] allowing for multiple ensemble NNs to be generated through training a single model. Ensembles comprise of periodic model *snapshots* taken during training. Diversity between each model snapshot is encouraged through specific learning rate (LR) scheduling between each snapshot. Specifically, a cyclic cosine function [192] repeating based on set training iterations:

$$\alpha(t) = \frac{\alpha_0}{2}\left(\cos\left(\frac{\pi \mod \left(t-1, \lceil \frac{T}{M} \rceil\right)}{\lceil \frac{T}{M} \rceil}\right)+1\right) \tag{5.2}$$

where the LR, α, is dictated by scaling the original LR, α_0, based off the current epoch t's position within the shifted sub-cosine function. Each of the M number of cosine functions are spread equally along to the total epoch count, T.

The resulting LR progression over a cosine cycle resembles a rapidly descending LR from an initial large value, gradually reducing in gradient to a set iteration and an assumed model convergence at local minima. At which point, model parameters are saved as a single ensemble snapshot before a large spike in LR is introduced to repeat the cosine cycle. Said LR spike

"dislodges" the model from the local minima allowing for descent into a potentially new local minima and resulting new unique ensemble model.

The resulting unique snapshot sub-models form a large combined final model for use in the testing stage. Final predictions are formed from the combined average of each predicted output probability of each snapshot model. The aggregation based on predicted probability as opposed to a voting style aggregation approach such as in random forest (RF) enables the weighting of model confidence within the final prediction result. Each snapshot NN within the overall ECNN architecture consists of a 2 hidden layer architecture containing 50 and 30 perceptrons accordingly. Perceptron counts were chosen using a simplistic grid search hyper-parameter optimization algorithm to provide best model performance.

The result of which, as indicated by Huang *et al.* , provides superior model accuracy and generalizability with similar training durations as compared to traditional momentum based learning rate schedulers. Such behaviour additionally provides potential to encourage divergent sparse first layer weights in combination with the aforementioned entropy weight regularization (See fig. 5.4). The result of which, provides diverse feature combinations for analysis.

5.2.4 Feature Ranking & Selection

Features can be categorized based upon the sparse weight matrix into three categories as detailed in section 5.2.2. An evaluation metric was designed as shown in Equation (5.3) called *Feature Sparsity Importance* to provide the capability to rank and identify possible features. Overall, feature ranking is based off the perceptron weight parameters directly associated to each feature between the input and first hidden layer of each snapshot using the following equation:

$$R_k = \frac{\overline{|W_k|} - \sigma^2(|W_k|)}{\max(\overline{|W_k|})} \tag{5.3}$$

where $\overline{|W_k^|}$ is the mean absolute weight on a column by column basis representing the mean weight associated with feature k. A higher mean absolute weight will generally indicate a feature of higher importance. In order to account for element-wise sparsity within the weight matrix, the variance of the absolute weights, $\sigma^2(|W_k|)$, is also taken into account:

$$\sigma^2(|W_k|) = \frac{\sum_j^J (w_{jk} - \overline{w}_k)^2}{J-1} \tag{5.4}$$

where a high value indicates high element-wise sparsity and vice versa. The maximum mean absolute weight used within the denominator ensures a non-dimensional value normalised to

$\{0,1\}$. The feature sparsity importance metric will evaluate fully connected features with high mean and low variance highly, partially connected features with high mean and high variance lower and finally disconnected features of low mean and low variance to a value near zero indicating low overall importance to the predictive model. Feature importance values from each snapshot model were averaged to obtain the final Feature Sparsity Importance value for use in thresholding.

Feature thresholding can be performed using various schema. Methodologies such as selecting based off a 95% importance cut-off would provide an effective adaptive threshold emphasising predictive performance. Such a cut-off would however produce a 107 feature subset, whilst a significant reduction, would still remain cumbersome in an application standpoint. A simple top $k = 10$ cutoff threshold provides a rather naive threshold policy, however coincidentally, as shown in fig. 5.2, a normal distribution fitted across a feature importance histogram highlights the predominance of low importance features whilst 10 features lie high outside the three standard deviation range. As such, these features are selected as the subset for further analysis.

Such feature ranking within the original data space contrasts highly with traditional statistical modelling techniques such as PCA or linear discriminant analysis (LDA) requiring orthogonal transformation into an embedding space for dimension ranking. As such, ECNN enables a direct interpretable ranking of individual medical events as predictive indicators of future hospitalization.

5.3 Experiment

The dataset population was extracted through the SAIL data-bank which consists of linked and coded patient records catalogued from various primary and secondary health services provided by the Welsh NHS, UK. Accordingly, data coverage encompasses the majority of the Welsh population, a total of 3 million individuals[193].

The Primary Care GP dataset (GP) contains individual medical records obtained from the various primary care practices around Wales. Every individual contains timestamped records of various events such as: prescribed medication, lab test results, and diagnoses coded as NHS read codes. The Patient Episode Database for Wales (PEDW) dataset comprises of attendance and clinical information for all hospital admissions within Wales. A continuous period of treatment for an individual can be traced from entry, to diagnosis, to hospital transfer (if any), to

treatment, to discharge. Information such as date of birth, gender, area of residence, deprivation score, etc. are provided if available for both datasets.

Table 5.1: Table containing read codes associated with a positive dementia diagnosis.

Read Codes					
E00..	E003.	Eu001	Eu021	E000.	E004.
Eu002	Eu022	E001.	E0040	Eu00z	Eu023
E0010	E0041	Eu01.	Eu024	E0011	E0042
Eu010	F11x2	E0012	E0043	Eu011	F11x5
E0013	E004z	Eu012	F11x6	E001z	E012.
Eu013	F11x7	E002.	E0120	Eu01y	F11x8
E0020	E041.	Eu01z	F11x9	E0021	Eu00.
Eu02.	F11xz	E002z	Eu000	Eu020	Fyu30

Data preparation involved the selection of all patients with a positive diagnosis of dementia based upon NHS read codes as indicated in Table 5.1[194]. Of note is the hierarchical nature of read codes which allows for a general broad consolidation of dementia diagnosis for simplification. Such examples include codes such as 'E00..' indicating all variations of code values possible on positions containing the decimal point. In practice however, there is inconsistent inclusion of both categorical and sub-categorical read codes within the dataset. As such, all categorical and sub-categorical read codes for dementia were included to ensure thorough consideration of all indicated dementia patients.

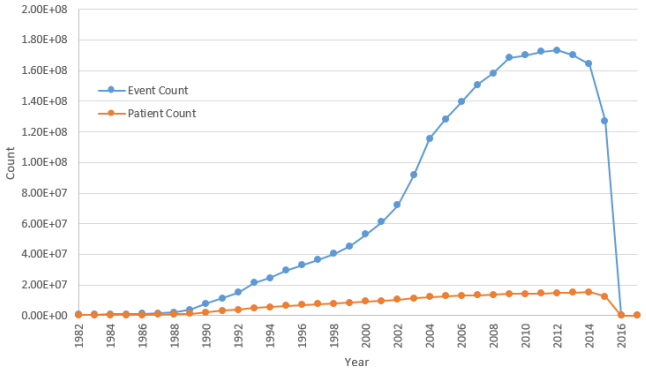

Figure 5.1: Graph indicating distribution of patient and event counts aggregated by year across the GP and PEDW datasets used for evaluation. As shown, the majority of patients and events span across a timeframe between 1982 to 2015. Of note, is the non-linear correlation between patient count and event count highlighting an increased frequency of recorded events over the years.

The overall dataset consists of the medical history from 1908 to 2017. However, dataset distribution by year as shown in fig. 5.1, indicates the vast majority of patient events distributed between 1982 to 2015. As such, patients and corresponding records have been limited to

the aforementioned time window. The selected population results in a gender split of 34.9% male and an overall mean age of 91.8 and a 10.5 standard deviation. The generally older population characteristic of our dataset provides opportunity for analysis into an especially more vulnerable age range of the general population more prevalent to dementia and resulting hospitalization or institutionalization. Further statistical population characteristics are shown in table 5.2.

The resulting dataset consists of 59,298 patients diagnosed with dementia. Patient timelines were selected one year before dementia diagnosis up to hospital admission if at all. An individual patient history, or sample within the input dataset, consists of a frequency table counting number of times specific medical events occur in a one year lead up to first hospitalization event. With a formatting similar to traditional one-hot encoding, the feature set comprises of all possible unique medical events which have occurred within the considered population resulting in 54,649 unique features or event codes. Whilst, the sum total frequency of all occurring medical events with in the population totals 52.5 million events, with a single individual medical history only using a small subset of said unique events, a significantly sparse dataset is produced, effectively highlighting the challenging extent of high dimensionality and data sparsity inherent within patient medical histories constructed into datasets for ML modelling. Consequently, such dataset properties provide an excellent opportunity for verification of ECNN.

Table 5.2: Statistical characteristics of sampled population

Category	Gender	Cond. +ve	Cond. -ve	Total
Mean Age	Male	85.18±9.44	93.62±10.56	89.04±10.82
	Female	88.98±8.51	97.41±9.62	93.27±10.02
	Total	87.56±9.05	96.19±10.09	91.80±10.50
Mean Event Count / Person	Male	1256±1127	443±565	885±1000
	Female	1318±1220	469±624	886±1053
	Total	1295±1187	461±605	872±884
Population	Male	11233	9441	20674
	Female	18945	19679	38624
	Total	30178	29120	59298

Condition positive and negative within the table indicates the sub-set of population with dementia having had a hospitalisation event after dementia diagnosis (+ve) or no recorded hospitalisation event (-ve).

As mentioned previously, the evaluation criteria for our methodology will be in predicting whether a dementia patient stays within a GP setting with minor accidents and events (condition negative) or whether a patient is admitted into a hospital setting due to major accidents or continued degradation of mental ability (condition positive). This will be indicated through a

lack of hospital data throughout a patient's time-line. The resulting patient dataset split consists of 30,178 patients admitted to hospital and 29,120 patients which remained within a GP setting.

A comparative evaluation between a similar traditional classification model with capability for feature ranking, RF was performed using the exact same dataset. Feature ranking on RF was produced through the use of traditional out-of-bag error comparison to perturbed datasets[195]. Additional comparative evaluation was also performed with a baseline methodology through a subset of 10 random features selected amongst the original overall feature-set via random number generator. Said random feature selection count of 10 coincides with selected feature count within the proposed methodology and in RF to enable direct comparison. Random selection of features is applied to highlight a baseline predictive capability of 10 features within the dataset.

5.4 Results

Experimental evaluation can be categorized into three distinct categories: predictive performance using the full dataset (section 5.4.1), analysis of model characteristics to produce a feature ranking (section 5.4.2), and final evaluation of feature ranking and selection against baseline methods (section 5.4.3). All experimentation was cross-validated using a 5 fold, traditional k-fold validation paradigm. In which, three folds are designated as the training set, one for validation and one for final testing in a cyclic sequence; repeated twice over. The resulting 5×2 test fold sequences of results are aggregated and presented within the remainder of this section.

5.4.1 Full Feature Results

The performance of ECNN as a pure classification model was assessed on the full set of features in comparison to a traditional classification methodology with combined feature ranking capability, RF. The intuition of such an assessment, in combination with section 5.4.3, being the evaluation of the validity of resulting feature rankings from ECNN.

Results are presented in table 5.3 showing aggregated predictive performance across various metrics with T-test to distinguish significance between the two methodologies. As shown, ECNN provides significant improvements (< 0.05 P-value), around 5%, in true negative rate (TNR) and positive predictive value (PPV) compared to RF whilst maintaining insignificantly near similar performance in true positive rate (TPR) and negative predictive value (NPV) re-

Table 5.3: Full feature set classification results.

Metric	ECNN Mean±Std. Dev.	ECNN 95% CI		RF Mean±Std. Dev.	RF 95% CI		P-Value
True Positive Rate	**0.746±0.036**	0.719	0.773	**0.746±0.005**	0.742	0.750	0.986
True Negative Rate	**0.762±0.043**	0.729	0.794	0.714±0.007	0.709	0.718	0.004
Positive Predictive Value	**0.766±0.024**	0.748	0.785	0.710±0.005	0.706	0.713	2.27E-06
Negative Predictive Value	0.744±0.019	0.730	0.758	**0.750±0.003**	0.747	0.752	0.404
Accuracy	**0.755±0.005**	0.750	0.757	0.729±0.002	0.728	0.731	2.61E-11

Comparative analysis of predictive performance between ECNN and random forest - a traditional classification model with the capability to perform feature ranking and selection. As shown, ECNN provides statistically significant, superior accuracy whilst providing superior feature selection (see table 5.7).

sulting in an overall superior model performance in accuracy. A major consideration however, is the larger variation in predictive performance of ECNN as compared to RF. Such variation was found during testing to be caused in part from the use of entropy regularization settling into perhaps a sub-par local minima of sparse weights producing inferior performing model snapshots affecting overall stability during the final prediction aggregation of the ensemble models.

The resulting overall performance improvement over RF however, comes with a major compromise of training complexity and duration as is standard in a comparison of RF to NN trade-offs. With a significant difference between RF and ECNN of 44 seconds to 2 hours average training duration respectively, such vast differences highlights the greatest disadvantage of ECNN and deep NN complexity overall. However, with a significant improvement in both predictive performance and feature ranking capability, as shown in section 5.4.3, such performance may justify the differences in training times.

5.4.2 Feature Selection

Within this section, we will present and analyse the resulting ensemble snapshots using the aforementioned feature ranking metric presented in section 5.2.4.

As shown in fig. 5.2, entropy regularization was able to successfully separate the majority of layer weights into a sparse filter mapping of values close to zero and one. Figure 5.4 alternatively provides a heatmap representation of the sparse first layer weights of each snapshot ensemble model produced. As seen, each ensemble mostly resembles each other with subtle differences highlighted in fig. 5.4 showing normalized difference of first layer weights between each pair of ensemble models. As such, snapshot ensembles are shown to successfully dislodge settled weights to generate new feature maps. Of note is how weight variance

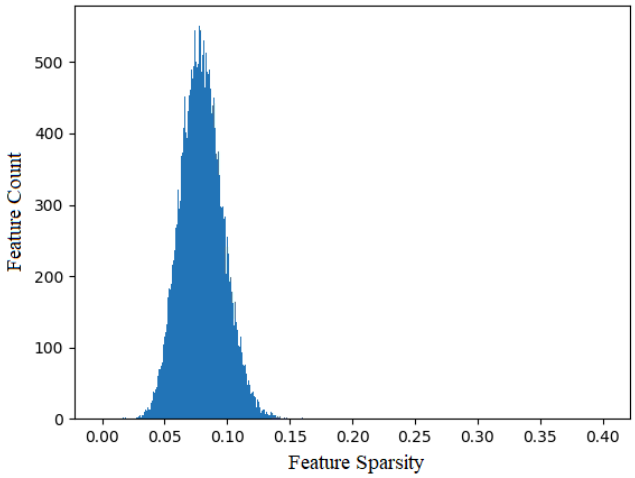

Figure 5.2: Histogram of features over mean importance factor across all snapshot ensembles of a randomly selected cross-validation run. As seen, the majority of features are normally distributed ($\mu = 0.0777, \sigma = 0.0265$) around a low overall feature sparsity value, indicating the majority of features introduced to ECNN are of low importance in prediction of hospitalization. Unable to be effectively shown, due to graph scaling constraints, 10 features lie outside 3 standard deviations of the distribution, shown in table 5.4.

between ensembles centres around specific features; as opposed to across layer 2 nodes or a combination of both. Consequently, such behaviour can be interpreted as high feature variance between ensembles indicating uncertainty of feature importance whilst low variance indicates a convergence of such features into a stable configuration of importance.

The proposed feature ranking metric was applied to the first layer weights of each ensemble and aggregated into a single normalized feature importance value for each individual feature. Figure 5.2 indicates the distribution of features across the feature importance spectrum. As seen, the majority of features form a normal distribution low on the feature importance metric with mean, $\mu = 0.0777$, and standard deviation, $\sigma = 0.0265$; whilst several features lie high on feature importance outwith the normal distribution by greater than three standard deviations. Consequently, these 10 outlier features were selected as the subset of important features used for continued further analysis, in addition to subset predictive performance testing in section 5.4.3.

These 10 medical events, summarized in table 5.4, form a varied collection of medical diagnoses, medication prescriptions and procedural events. Qualitative analysis and literature review of the identified medical events show effective feature selection from ECNN with every event occurrence being either positively associated to an increased hospitalization risk or

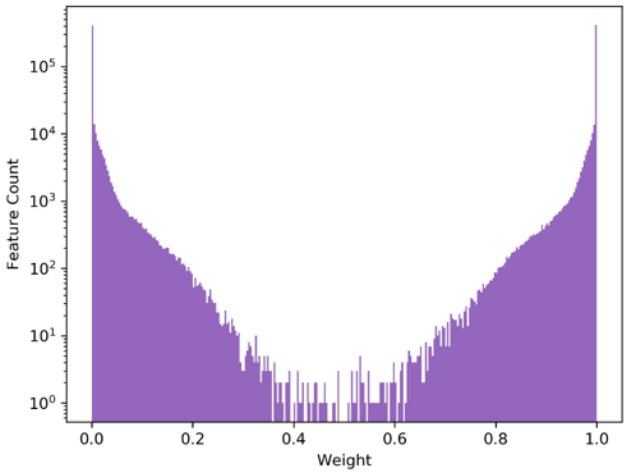

Figure 5.3: Shown, is the log scaled histogram of final model weights of the first layer of a randomly selected model within cross-validation. As seen, the vast majority of weights have converged to values close to $\{0, 1\}$ in response to the proposed entropy weight regularization. As mentioned, a comparatively small set of weights (an order of magnitude less than successfully separated) show a non-perfect separation towards either extreme. Further analysis of said weights indicate belonging to specific features, contributing to the ultimate variance between each ensemble, as highlighted in fig. 5.4.

present an entirely novel or inconclusive association.

In regards to established direct risk factors identified by ECNN, a literature review is presented highlighting each positive correlation. As shown, a diagnosis of essential hypertension or idiopathic hypertension was identified as the highest ranked feature with an average importance factor of 0.481, vastly exceeding the exhibited normal feature distribution mentioned previously. Of course, such a correlation between hypertension and hospitalization incidence has already been shown to exist through cohort studies[196, 197]. Previous literature have also studied several other risk factors identified by ECNN. In regards to the second most highly ranked event, prescription of *Adcal-D3 - calcium and vitamin D* supplements, under the assumption of a resulting vitamin D or calcium deficiency in the individual, studies have shown general increase in hospitalization risk for the elderly from resulting co-morbidities[198] in addition to direct potential risk[199, 200]. *Influvac*, a flu vaccine, the third highest ranked event, regularly prescribed to highly at risk elderly individuals, highlights established risk factors of influenza on functional decline within the elderly[201]. Additionally, blood glucose lab tests for potential diabetes and *simvastatin*, prescribed for high blood cholesterol are further established risk factors for general hospitalization risk in the elderly demented population[197]. Osteoarthritis, a condition with a common prescription of *Ibugel*[202] - a gel based ibuprofen

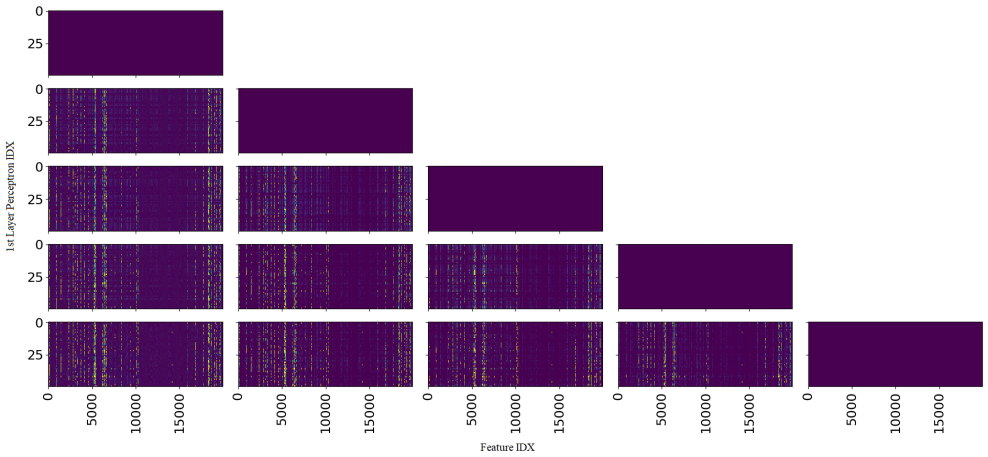

Figure 5.4: Shown, is a complete comparative heat map matrix of the absolute differences in weights between the first hidden layer of every possible pair of the 5 produced snapshot ensembles. Colour values are mapped so that yellow hot represents values close to 1 whilst purple cold represent values close to 0. Left-top to right-bottom diagonals show weight difference between the same ensemble and are thus irrelevant for analysis. Of note, the vertical patterning for each heat map indicates any weight differences between snapshot pairs are focused on specific individual features across all snapshot pairs. This suggests a convergence in importance factor for the vast majority of features with a small but consistent subset of edge-case features producing variation amongst the snapshot ensembles.

medication identified as 7th on the list, is also widely regarded as a hospitalization risk factor of the elderly[203]. Such feature to class correlation can be the result of prior bias to such an elderly data population.

Prescription of *Serc-16 tablets*, prescribed for Ménière's disease, presents an interesting secondary indicator of hospitalization risk. With symptoms of vertigo, titinnus, and hearing loss - Ménière's disease associates with increased fall risk in the elderly[129] resulting in indirect risk of hospitalization.

As shown, the identification of already established risk factors by ECNN demonstrates effective risk factor recognition, highlighting the potential for further clinical analysis on the remaining medical events for potential correlations. Of the remaining event indicators: *Social group 3 - skilled*, occurrence of nightmares and encounter between GP and a third party in regards to the patient; little or inconclusive studies have attributed such events as a precursor to hospitalization. Van de Vorst *et al.* indicates no statistical significance for hospitalization risk between mid-tier socioeconomic status, generally associated with a skilled individual, and high or low-tier status. There was however, positive significant correlation from low to high-tier status[204]. Nightmares have potential to be associated with symptoms of delirium, the

result of which, hospitalization risk is increased[205]; however, such a generic medical event with multiple associations to various conditions would require further study to be presented as an indicator on it's own. Finally, third party encounter addresses a wide range of situations involving reports by individuals related to the individual suffering from dementia. Whilst it has been established that dementia detection is predominantly reliant on self-reporting or by relatives[108], no literature was found studying hospitalization resulting from non-emergency third party reports.

Linear independent correlations between the identified medical events to hospitalization incidence was analysed through Pearson's correlation and reported in table 5.4. Interestingly, there seems to be little correspondence between r value and ECNN ranking and in some cases, little statistical significance. Such behaviours indicate a distinct lack of independent linear correlations on individual risk factors. Tests on modelling hospitalization prediction using NN and RF on the individual, identified features provide no discriminative capability; requiring all 10 features to produce predictive performance indicated in section 5.4.3. Such observation hints at the capability of the underlying NN architecture of ECNN being able to formulate non-linear relationships between features, consequently being unable to produce individually discriminative medical events. The extraction and interpretation of non-linear combinatorial relationships between features remains an open avenue for further research of great benefit within the medical informatics field.

Table 5.4: Top 10 event codes ranked in order of importance as determined by ECNN.

Importance	Event CD	Definition	Clinical Indication	Pearson's r	P-Value
0.481	G20..	Essential Hypertension	Essential Hypertension	0.1504	4.65E-297
0.318	ip3j.	Adcal-D3 1.5g/10ug chewable tablet	Vitamin D & Calcium Deficiency	0.0695	1.98E-64
0.300	n473.	Influvac sub-unit prefilled syringe 0.5mL	Annual immunisation against flu	0.0635	4.44E-54
0.259	44Uz.	Blood glucose raised NOS	Type-2 Diabetes	0.0223	5.71E-08
0.254	bxd5.	Simvastatin 40mg tablet	Hypercholesterolaemia	0.1879	0.0
0.247	1323.	Social group 3 - skilled	-	0.0040	0.326
0.234	dh12.	Serc-16 Tablet	Ménière's Disease	0.0107	9.39E-03
0.234	ja1I.	Ibugel gel 100g	Osteoarthritis	0.0154	1.71E-04
0.231	E2749	Nightmares	-	0.0070	0.087
0.227	9N32.	Third Party Encounter	-	0.0626	1.23E-52

A Pearson's r statistic was used to measure linear correlation between medical event occurrence and hospitalization incidence in dementia patients. As shown, most of the ranked events show statistically significant linear positive correlation to hospitalization incidence highlighting potential predictors of hospitalization events post indication of such medical events. Of, note is the ranking of event codes based off r statistic being non-similar to the ranking produced by ECNN. Of course, with r statistic analysing only linear bivariate correlations, any non-linear multi-variate associations between event and hospitalization incidence detected by ECNN are lost, explaining the discrepancy between r statistic ranking and ECNN ranking. Multi-variate statistical analysis is further warranted on the identified events is a potential avenue of further investigation.

Table 5.5: Top 10 event codes ranked in order of importance as determined by random forest.

Importance	Event CD	Definition
0.0164	44M3.	Serum total protein
0.0162	42L..	Basophil count
0.0136	44D6.	Liver function test
0.0112	44M4.	Serum albumin
0.0100	451E.	GFR calculated abbreviated MDRD
0.0085	44P5.	Serum HDL cholesterol level
0.0083	44I5.	Serum sodium
0.0083	44I8.	Serum calcium
0.0080	1Z12.	Chronic kidney disease stage 3
0.0071	9N36.	Letter from specialist

Whilst RF highlighted entirely separate features of importance compared to ECNN, several event codes can be hypothesised to indicate similar clinical significance. For instance, serum calcium indicates the use of calcium level blood tests potentially related to the dispensing of Adcal-D3 tablets as indicated by ECNN. Similar indications can be found with cholesterol level and the dispensing of Simvastatin. Interestingly, a high focus on blood work is shown with 8 of the top 10 event codes shown being lab results of blood work.

5.4.3 Reduced Feature-set Predictive Performance

Several comparisons were evaluated to determine feature selection performance. The reduced subset of features produced by ECNN were used to train on various standard classification methodologies as a comparison to the full dataset. The top 10 features ranked by RF, shown in table 5.5, were also used as a baseline comparison of a traditional effective feature selection procedure while a random 10 feature selection was also evaluated to provide an indicator of dataset baseline predictability. The results are shown in table 5.7.

Table 5.6: Logistic regression comparison using reduced feature selection results.

Metric	**ECNN Features**			**Full Features**			**P-Value**
	Mean±Std. Dev.	95% CI		Mean±Std. Dev.	95% CI		
TPR	**0.685±0.008**	0.679	0.691	0.286±0.035	0.260	0.313	4.02E-16
TNR	0.711±0.009	0.705	0.718	**0.804±0.016**	0.792	0.816	0.014
PPV	**0.746±0.010**	0.738	0.754	0.487±0.032	0.463	0.511	4.63E-14
NPV	**0.646±0.007**	0.640	0.651	0.633±0.024	0.615	0.651	1.84E-09
Accuracy	**0.697±0.005**	0.693	0.701	0.600±0.018	0.586	0.613	2.82E-15
AUROC	**0.719±0.007**	0.717	0.721	0.386±0.034	0.376	0.397	2.68E-61

Shown is a comparative analysis of two logistic regression classification models, one trained on ECNN feature selections and one on the full set of dataset features. As shown, ECNN feature selections result in superior classification performance whilst providing a significant reduction in feature size.

In a direct pair-wise comparison of predictive performance of feature ranking based on ECNN versus RF for each of the baseline NN and RF classification models shows generally superior performance using features ranked by ECNN. This is highlighted by a 4.1% improvement in F1 score between RF and proposed when using a NN baseline classification model, and a 1.4% improvement using a RF baseline classification model.

Table 5.7: ECNN and random forest reduced feature selection results.

Metric	ECNN Features		RF Features				Random Selection Features	
	NN	RF	NN	P-Value	RF	P-Value	NN	RF
TPR	0.758±0.014	0.721±0.0008	0.775±0.039	0.077	0.720±0.003	0.125	**0.998±0.002**	0.688±0.540
TNR	0.759±0.025	**0.780±0.0012**	0.659±0.056	2.76E-11	0.743±0.002	2.00E-79	0.002±0.002	0.323±0.559
PPV	0.766±0.016	**0.773±0.0008**	0.704±0.024	0.06E-19	0.745±0.001	1.21E-90	0.510±0.001	0.566±0.098
NPV	**0.751±0.005**	0.728±0.0004	0.739±0.017	0.002	0.719±0.002	1.42E-36	0.308±0.314	0.233±0.251
FPR	0.241±0.025	**0.220±0.0011**	0.341±0.056	2.10E-12	0.257±0.002	1.50E-78	0.998±0.002	0.677±0.559
FNR	0.242±0.014	0.280±0.0008	**0.225±0.039**	0.062	0.280±0.003	0.490	0.002±0.002	0.312±0.540
Accuracy	**0.759±0.006**	0.750±0.0003	0.718±0.007	8.71E-36	0.732±0.001	1.70E-79	0.510±0.001	0.509±0.001
F1 score	**0.762±0.002**	0.746±0.0003	0.737±0.005	2.15E-38	0.732±0.001	3.82E-73	0.675±0.001	0.489±0.322
AUROC	**0.854±0.015**	0.785±0.0008	0.734±0.031	4.54E-30	0.735±0.002	6.32E-91	0.591±0.002	0.604±0.532

Shown is a comparison of feature selection effectiveness between ECNN, random forest - a similar traditional feature selection methodology, and random feature selection. Two types of basic classification models: neural network and random forest, trained on the various feature selection subsets, are used in determining selection effectiveness via predictive performance. A baseline random subset of features was also evaluated P-value analysis is a comparison of ECNN features against RF features of the corresponding underlying predictive model.

A definitive superior baseline model in an application standpoint for our feature subset use case however, is not as clear cut; with RF providing superior TNR with comparable accuracy scores and NN providing overall best F1 score and accuracy. In consideration of an application based hospitalization warning system, NN provides the superior NPV and as such, the superior screening type test for high risk demented patients.

In regards to the baseline random feature selection process, both feature selection methodologies produced results significantly improved over that of random guessing. Of note however, is the inability of NN in training an effective classification model when using the randomly selected feature subset, with final inactive models producing continuous positive predictions resulting in a 'superior' TPR. Additionally, RF also produced generally inactive models using the random feature subset, swinging between continuous positive or continuous negative predictions indicated by significantly large standard deviations. As such, random feature subset results do not provide an effective comparison of proportional predictive performance as compared to non-random feature selection methodologies but instead highlight the difficulties of selecting small subsets of features able to adequately model patient hospitalization.

In reference to table 5.6, feature ranking and selection using ECNN shows a statistically significant improvement in overall predictive performance as opposed to the use of the full feature dataset using a traditional logistic regression classification model. Said results highlight the challenges of such a high-dimensional and sparse dataset and the advantages of effective feature selection, enabling effective modelling of the problem space in a significantly reduced subset of features. Such complexity reduction is emphasised in the contrast of average training durations with logistic regression trained on the full set of features requiring 33 minutes whilst

training on a subset of 10 features requiring seconds.

5.5 Conclusion

This chapter proposes a novel combination of methodologies for the prediction of hospitalisation potential with patients suffering from dementia. Using a novel adaption of snapshot ensembles to use a dynamically generated learning rate schedule, in addition to an adaption of entropy weight regularization for use with NNs and subsequent novel evaluation of model parameters: we were able to identify 10 medical events highly indicative of future hospitalization of demented individuals out of an extremely high dimensional and sparse dataset of 54,647 unique medical events. Comprising of diagnostic events, medication prescriptions and procedures, said events were able to model and predict future hospitalization to a performance equal (and in certain cases better) than that of the full dataset. ECNN provides significant advantages to statistical feature selection methods in interpretability, and additionally in ML based modelling techniques in predictive performance.

The identification of said medical events opens avenues for the potential creation of early warning systems to identify demented individuals at high risk of hospitalization or institutionalization. With multiple indications of nutritional health being a major impact in hospitalization risk factor, such information can be further investigated for potential prevention through an emphasis in improved nutritional care for dementia patients. Such examples highlight the many possibilities focusing on pre-empting and preventing hospitalization through alteration of secondary care practices. Overall contributions such as those indicated allow for a potential reduction in critical healthcare utilization, itself a positive advancement, whilst reducing risk in a statistically elderly and vulnerable population through reduction in exposure to hospital induced risk factors such as infection.

Multiple aspects of the ECNN algorithm presented within this chapter are open for improvement. The application of longitudinal health data is not fully exploited using the ECNN algorithm as a result of a non-recurrent based prediction model based off traditional perceptron NNs. Consequently, major considerations of longitudinal health data such as: the continually changing health picture of an individual, the degenerative change in mental capability due to dementia, and other such longitudinal aspects are not fully modelled or analysed. Arguments can be made on such longitudinal information being learned through optimization within the feature representation composed by a considerably large enough capacity model such as a time-distributed convolutional neural network (CNN); however, such architectures suffer from

issues of over-fitting due to such high learning capacity in lieu of regularization driving towards such a "recurrent" feature representation. Further complications of finite impulse response limitations within CNNs due to predefined kernel dimensions compared to potentially infinite impulse response from recurrent neural networks (RNNs) (or more accurately, long short-term memorys (LSTMs) if vanishing gradient is considered) limit the overall effectiveness of the traditional NN in longitudinal applications. Following chapters such as the time-series LSTM based model in chapter 7 seek to leverage the temporal aspect of health records to a greater degree than as presented within this chapter.

The collection of medical events highlight already established risk factors for hospitalization indicating effective capability whilst novel events present opportunity for further focused traditional clinical analysis as potential risk factors and indicators. As such, ECNN provides future potential for use within other medical informatics domains as risk factor identification. The general nature of patient medical records, in conjunction with ECNN enables application within other domains to provide interpretable, small-scale indicators allowing for ease of identification of at risk individuals for pre-emptive care.

Of significant interest, the corresponding medical events highlighted within our feature selection approach presented in this chapter highlight already established risk factors known within the literature. Such medical events indicate effective and relevant selection by the proposed architecture, in-line with current medical understanding. Additionally, clinical indications such as hypertension, diabetes, and hypercholesterolaemia remain significant in further independent feature selection based studies and chapters within this thesis, indicating prevailing trends of significant biomarkers.

Chapter 6

Modelling Severe Sepsis Onset: Boosted Cascading LSTMs

Contents

6.1 Introduction

Within the previous chapter, we presented a novel embedded feature selection approach for the modelling of long-term dementia related hospitalisation events. The lack of temporal based feedback as a patient's health changes across time highlights a major weakness within the aforementioned modelling approach. Proceeding onwards, we move towards the use of recurrent based modelling approaches, able to exploit the longitudinal relationships of the continually progressing health of an individual.

Within this chapter, we present the opposing characterised clinical objective of sepsis prediction within critical care settings such as the intensive care unit (ICU). Section 4.3 highlights the severe negative clinical outcomes of severe sepsis and septic shock even within the domain of modern sterile clinical medicine. Modern sepsis detection remains reliant on human vigilance with reference to simple, rules-based multi-categorical scoring of vital-organ health as indication of potential sepsis. There exists a unique opportunity to incorporate machine learning (ML) based applications into hands-off continuous monitoring platforms to predict sepsis development in a timely manner at greater prediction accuracy than the currently established diagnosis systems. As previously presented, mortality rate increases of 5-8% per hour of undiagnosed and untreated sepsis[171, 176], highlights the importance of early detection and decisive treatment of sepsis.

As highlighted in chapter 4, ICU based health data characteristics runs counter to that of long-term general practice (GP) records used within the previous chapter. The latter—consisting of long-term, variable-frequency data of significant sparsity—highly contrasts the data characteristics of the former; being short-term, high-frequency data consisting of comparatively smaller data dimensionality requiring time-critical response. Of particular consideration for application of sepsis prediction is the comparatively low prevalence of sepsis positive events over the period of a critical care patient timeline.

Such low prevalence of sepsis event indications results in highly imbalanced positive/negative sepsis class distributions (less than 1% positive time-steps in many study datasets presented within this thesis) in combination with attributed big data issues of data sparsity and high-dimensionality results in a highly non-trivial modelling challenge. Consequently, this chapter presents a novel cascading ensemble deep learning (DL) based approach to time-series prediction with the aim of addressing the unique aforementioned challenges of detection of a low prevalence medical condition across a large population. The proposed methodology, using ICU monitored patient vital data, outperforms current state-of-the-art methodologies within

literature for sepsis prediction, six hours prior to current detection times. The combined use of a novel cascading approach and loss objective alleviates inherent issues within medical informatics based datasets of high data sparsity, high dimensionality and severe class imbalance. Through evaluation on several ICU based datasets, we demonstrate superior performance and model generalisability.

6.2 Dataset

Experimental evaluation will be applied via two unique datasets comprised of time-series patient medical records within an ICU based setting. The objective for evaluation, as mentioned previously, will be the accurate prediction of sepsis onset within a septic patient, 6 hours prior to clinical diagnosis by a human medical practitioner. The aforementioned datasets will be the Physionet Computing in Cardiology Challenge 2019 dataset (CinC 2019)[206] and the Medical Information Mart for Intensive Care dataset (MIMIC)[57]; both publicly available open-source large-scale datasets.

Of reviewed similar studies, 7 of the 11 directly related papers feature the publicly available MIMIC dataset as a primary or secondary dataset. Being an open-source dataset of real-world deidentified ICU medical data of over 40,000 patients, MIMIC gains much popularity due to the scarcity of openly available datasets of such size within the medical informatics domain. With patient data privacy a continual concern within medical informatics, there remains a shortage and need for diverse open datasets of such nature whilst maintaining said data privacy.

In regards to dataset characteristics, all studies follow similar patterns of using binned, hourly sampled features consisting of patient vital signs (heart rate, oxygen saturation, blood pressure, etc.), laboratory results (lactic acid, platelet count, urine output, etc.) and patient demographics (age, sex, race, etc.). Prediction objectives mainly focus around the prediction of sepsis onset 3-4 hours earlier than that of recorded medical suspicion. Considered population sizes reported within studies range drastically from 140 to 32,000 patients with data collection periods spanning back to 2001.

6.2.1 Computers in Cardiology 2019

The CinC 2019 dataset comprises of ICU patient medical records from two separate hospital units spanning the years 2010 to 2020. The 40 unique features provide hourly patient level data consisting of continuous vital signs, sparse lab test results and static demographic in-

formation. In regards to data sparsity, the various feature categories present varied levels of missing data. Vital sign category features average 32.4% missing data in which otherwise regular hourly measurements are not recorded. In regards to lab test results, an average of 94.9% missing data is indicated. Such significant missing data is mainly due to lab tests only being irregularly ordered when required which, in combination with hourly binned timesteps, results in unavoidable significant data sparsity. Static demographic information contains no missing data.

Overall population demographics show a mean age of 61.6 years and standard deviation of 16.5 years. Gender proportions shift towards male with 55.9% proportion of males to 44.1% females. Sepsis prevalence within the CinC 2019 dataset shows 2932 (7.3%) patients having a positive sepsis event out of the total 40,336 ICU patients.

The prediction objective class labels are defined on an hourly basis as a binary value indicating development of sepsis as positive (1), or no indication of sepsis, negative (0). Said sepsis development has been defined clinically by Reyna *et al.* [206] as a series of medical events indicating clinical suspicion of sepsis development. Clinical suspicion of sepsis is thus defined as either a two point deterioration in sequential organ failure assessment (SOFA) score, antibiotic administration within 72 hours of blood lab culture testing, or blood lab culture testing within 24 hours of antibiotic administration. The sepsis prediction objective, 6 hours prior to official clinical suspicion, results in a class label vector shifted 6 hours forward for all overall positive sepsis patients.

Traditional linear correlation using Pearson's coefficient was also performed against said sepsis label with remaining features as highlighted in table 6.2. Said correlations highlight a patient's length of stay within ICU as the overall most significant indicator of sepsis development. Such a correlation is well known, with many studies highlighting the significance of extended hospital stay resulting in the increased risk of hospital-acquired sepsis[207, 208]. Alternatively, such correlation can be attributed in reverse, with the development of sepsis resulting in an overall longer length of stay, as highlighted by many studies on sepsis[209, 210, 175, 211]. Interestingly, the standard indicators of sepsis as indicated by SOFA: respiratory system (FiO2), cardiovascular system (MAP), liver function (bilirubin), blood coagulation (platelets) and kidney function (creatinine) do not correlate highly within the CinC 2019 dataset. Such behaviour can be attributed to the assumption by Pearson's coefficient of independent linear correlations between features and label, highlighting the disadvantages of such traditional statistical methodology in favour of more complex non-linear based methodologies for modelling

Table 6.1: Data attributes and missing data percentage of the PhysioNet CinC 2019 challenge dataset

Attribute	Missing Data (%)	Attribute Details
Vital Signs		
HR	9.9	Heart rate (beats per minute)
O2Sat	13.1	Pulse oximetry (%)
Temp	66.2	Temperature (Deg C)
SBP	14.6	Systolic BP (mm Hg)
MAP	12.5	Mean arterial pressure (mm Hg)
DBP	31.3	Diastolic BP (mm Hg)
Resp	15.4	Respiration rate (breaths per minute)
EtCO2	96.3	End tidal carbon dioxide (mm Hg)
Laboratory Values		
BaseExcess	95.8	Measure of excess bicarbonate (mmol/L)
HCO3	95.8	Bicarbonate (mmol/L)
FiO2	91.7	Fraction of inspired oxygen (%)
pH	93.1	N/A
PaCO2	94.4	Partial pressure of carbon dioxide from arterial blood (mm Hg)
SaO2	96.5	Oxygen sat from arterial blood (%)
AST	98.4	Aspartate transaminase (IU/L)
BUN	93.1	Blood urea nitrogen (mg/dL)
Alkalinephos	98.4	Alkaline phosphatase (IU/L)
Calcium	94.1	(mg/dL)
Chloride	95.5	(mmol/L)
Creatinine	93.9	(mg/dL)
Bilirubin_direct	99.8	Bilirubin direct (mg/dL)
Glucose	82.9	Serum glucose (mg/dL)
Lactate	97.3	Lactic acid (mg/dL)
Magnesium	93.7	(mmol/dL)
Phosphate	96.0	(mg/dL)
Potassium	90.7	(mmol/L)
Bilirubin_total	98.5	Total bilirubin (mg/dL)
TroponinI	99.0	Troponin I (ng/mL)
Hct	91.1	Hematocrit (%)
Hgb	92.6	Hemoglobin (g/dL)
PTT	97.1	partial thromboplastin time (seconds)
WBC	93.6	Leukocyte count (count*10^3/μL)
Fibrinogen	99.3	(mg/dL)
Platelets	94.1	(count*10^3/μL)
Demographics		
Age	0.0	Years (100 for patients 90 or above)
Gender	0.0	Female (0) or Male (1)
Unit1	0.0	Admin identifier for ICU unit (MICU)
Unit2	0.0	Admin identifier for ICU unit (SICU)
HospAdmTime	0.0	Hours between hospital and ICU admit
ICULOS	0.0	ICU length-of-stay (hours)
Class Labels		
SepsisLabel	0.0	Positive sepsis (1) otherwise (0)

and feature discovery.

6.2.2 MIMIC-III

The MIMIC dataset closely resembles the CinC 2019 dataset as de-identified, time-stamped patient records admitted to ICU units over a period of 2001 to 2012[57]. Population demo-

Table 6.2: Pearson's Correlation coefficient of features in relation to sepsis labels on the PhysioNet CinC 2019 dataset.

Attribute	Attribute Correlation	Attribute	Attribute Correlation
ICULOS	0.133774	Hct	0.010177
EtCO2	0.047623	Fibrinogen	0.010153
pH	0.034250	O2Sat	0.010075
PaCO2	0.034141	Gender	0.009280
HR	0.033029	Unit1	0.007716
Resp	0.027568	Bilirubin_direct	0.007272
Lactate	0.027548	Platelets	0.006361
SaO2	0.022655	AST	0.006331
BUN	0.019364	Glucose	0.005130
Potassium	0.018152	df_index	0.003528
WBC	0.014905	FiO2	0.002236
Chloride	0.014835	MAP	0.001800
Bilirubin_total	0.014017	TroponinI	0.001610
Phosphate	0.013819	Temp	0.001025
Alkalinephos	0.013072	Age	0.000191
Magnesium	0.012718	BaseExcess	0.000179
HCO3	0.012602	DBP	-0.001441
Creatinine	0.012027	SBP	-0.007564
PTT	0.011664	HospAdmTime	-0.019052
Calcium	0.011213	Unit2	-0.024292
Hgb	0.010250	Patient	-0.027861

graphics further closely resemble CinC 2019 with a 63.8 year mean age and 17.4 year standard deviation, a 55.9% male to 44.1% female split, and 36.8% (19,680) of hospital admissions indicating a suspicion of sepsis across the 53,423 total admissions. All medical events are encoded with the widely used International Classification of diseases, Ninth Revision (ICD-9) coding system, resulting in a large feature count of 5386 unique medical events across a total 181,253,575 entries. Occurrence counts across each unique medical event highlight a significant proportion of overall event entries representing a small sub-sample of common unique medical events. As such, low-occurrence, irrelevant features were cut based on selecting important features across a cumulative entry count across unique medical events, from most to least occurrence, up to a count total of 99% of overall entries. As a result, 71.7% (3861) features were dropped producing a final feature count within our dataset of 1525 unique medical events.

Initial data analysis via Pearson's correlation show similar linear feature correlations to that of the CinC 2019 dataset, as shown in table 6.3. Of the 5386 unique features, most of the said 10 largest correlations between feature and sepsis label feature within the CinC 2019 dataset. Albeit with differing sorting positions and in some cases (MAP and Age) opposite trending correlations.

The final considered dataset for MIMIC presents similar patient demographics and objective applications when compared to the CinC 2019 dataset. However, the large increase in feature count presents an extreme discrepancy in both data size and sparsity as compared to CinC 2019. As such, both datasets provide differing unique aspects and challenges across the common objective of sepsis prediction via patient vitals within an ICU setting.

Table 6.3: Top 10 Pearson's Correlation coefficient of the 50 most common features in relation to sepsis labels on the MIMIC-III dataset

Attribute	Event Count	Attribute Correlation
HR Alarm	1,582,597	0.2264
SaO2 Alarm	1,523,293	0.2192
Resp	1,571,434	0.2113
SaO2	1,571,188	0.2055
DBP	1,047,176	-0.1877
MAP	1,039,973	-0.1872
SBP	1,050,257	-0.1867
Heart Rhythm	1,431,019	-0.1848
SpO2	1,523,216	-0.1839
Age	3,979,611	-0.1827

6.3 Methodology

The proposed methodology consists of a small, standard long short-term memory (LSTM) model augmented by several novel concepts aimed at alleviating non-trivial difficulties encountered in time-series based medical records. Namely, large-scale datasets containing thousands of highly sparse non-independent features: the "curse of dimensionality". The application of a novel boosted cascading sub-network model training optimization strategy, supported by a novel shifting-margin hinge loss function, provides effective reduction to over-fitting performance issues favouring the significantly more common negative class stemming from major dataset class imbalances.

6.3.1 Boosted Cascading Sub-networks

As mentioned previously, there exists a large class imbalance towards condition negative samples, a common occurrence within diagnosis based applications within the field of medical informatics. With a 1.7% proportion of positive timestep samples across the dataset, such a class imbalance poses a non-trivial issue. Traditional methodologies for class imbalance such as over-sampling and under-sampling result in limited effectiveness[212]. We propose a more

effective novel combined model architecture and training strategy for solving said imbalance issues using a combination cascade, boosting ensemble training algorithm.

Boosting, the identification of hard to classify samples and subsequent emphasis in future weak learner sub-models, leverages the inherent advantages of adaptive ensemble based modelling. Simultaneously, said sample weighting can be adapted towards producing a soft filter training approach within the cascading concept of our methodology. Let each cascade sub-model, $m = 0$, be trained in a traditional manner as indicated in eq. (6.1). The minimization of model loss L^m as dictated by traditional loss functions $L(y_i^t, \hat{y}_i^t)$, based off true sepsis label, y_i^t, for patient i at timestep t and model prediction $\hat{y}_i^t$. Additional standard model normalization parameters, p, such as L1 or L2 norm or, in our specific application, those introduced in sections 6.3.3 and 6.3.4 at each sub-model are included in L_p^m.

$$L^m = \arg\min_L \sum_{i=0}^{M} \left(w_i^m \sum_{t=0}^{T} L(y_i^t, \hat{y}_i^t) \right) + L_p^m \tag{6.1}$$

The adaptive weighting factor, w_i^m for each patient, dictates sample importance within training loss. Lacking any initial sample difficulty indications, initial weighting factors are defined to weight heavily towards, and thus emphasise, the minor class. The goal of which, is to produce the initial filtering cascade sub-model emphasising near perfect classification of the minor class samples at the cost of a low classification rate of the major class. Subsequent cascades can thus be trained with an adaptive boosting sample weighting factor, based off previous model performance, to emphasise the remaining incorrectly classified major class predictions.

Said sample weight updates are thus calculated using the following equation:

$$w_i^m = (1 - \lambda_w) w_i^{m-1} + \frac{\lambda_w}{T} \sum_{t=1}^{T} \left| y_i^t - \hat{y}_i^{t,m-1} \right| \tag{6.2}$$

Patient weighting factors are driven towards lesser or greater importance based off the patient timeline's, T, mean absolute error generated by the previous cascade prediction results, $\hat{y}_i^{t,m-1}$. Where, y_i^t is the true class label and $\lambda_w \in \{0 < \mathbb{N} < 1\}$, the user defined weight hyperparameter coefficient regulating the influence of the current cascade error towards the weighting factor. Consequently, said hyperparameter provides control over driving the weighting factor more rapidly towards a pure boosting strategy ($\lambda_w \to 1$) or emphasise the initial cascading approach with smaller updates ($\lambda_w \to 0$)

Overall cascade generation can be driven by traditional meta-training evaluation metrics. As such, our approach utilises a minimum delta loss improvement stopping strategy or until a maximum cascade count hyperparameter is reached.

Final prediction methodology follows the aforementioned cascade approach, with minor class predictions filtered out in each cascade as said cascades prediction whilst remaining samples are passed to the next cascade model. The final cascade minor class predictions are then taken as the final subset of predicted minor classes, whilst major class predictions are appended to the larger major class prediction subset.

With the application of multiple cascading sub-models within the training procedure, there presents an opportunity to adjust per-cascade model complexity. The objective of which is to gradually increase discriminative capability between the increasingly complex filtered class boundaries as we descend deeper through cascade filters. With such a concept in mind, we now regard our model architecture. The initial sub-model composes of two groups composed of a LSTM hidden layer, followed by batch normalisation and dropout layer with 25% dropout rate. LSTM counts for each hidden layer are 16 and 8 respectively. Finally, the output layer composes of a single LSTM node for a binary prediction output. Each subsequent cascade sub-model increases the LSTM hidden layer's node count by an additional 25% of the previous cascade. Consequently, at our sixth cascade sub-model, hidden layer node counts are 31 and 61 respectively. Hyperparameter optimisation was determined through greedy grid search.

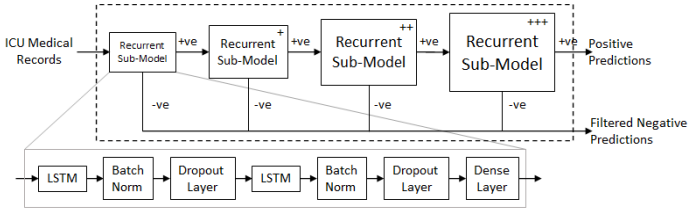

Figure 6.1: Diagram highlighting the general model architecture formed via the proposed boosted cascading sub-networks training methodology. As seen, each sub-model is formed of two hidden layer sets of LSTM, batch normalisation and dropout training layer followed by a final single perceptron output layer. LSTM counts for each set are initially set at 16 and 8 nodes respectively, increasing in count by 25% with each descending cascade. Dropout is set at 25% dropout rate. Overall architecture is formed of multiple sub-models arranged in a linear cascading manner with negative predictions filtered out and positive predictions passed onwards to each cascade until model end.

6.3.2 Shifting Margin Hinge Loss

As mentioned previously, within section 6.3.1, the proposed methodology applies increasingly larger capacity sub-models as cascades are formed. However, in consideration of the decreasing sample set size with the increasing model capacity, model over-fitting becomes a predominant issue during training. As such, we propose the novel shifting margin hinge loss function

to manage discrimination complexity across class boundaries, thus dampening potential over-fitting issues.

The shifting margin hinge loss is an adaption of the traditional hinge loss function and it's concept of forming the optimal decision boundary. As such, let the considered linear decision boundary be defined as the hyperplane, $x_i^T\beta+\beta_0$, defined by parameters β and β_0 where sample feature vectors are indicated by $x_i, i=1,\ldots,N$. Through the maximisation of distance, M, between decision boundary and opposing class boundaries, we facilitate the optimal separation of samples along said decision boundary. Class separation can thus be indicated by $y_i^+(x_i^T\beta+\beta_0)>0$ and $y_i^-(x_i^T\beta+\beta_0)<0$ where y_i^+ and y_i^- indicate positive and negative class samples respectively.

Within a simplistic linearly separable dataset, there exists infinitely many combinations of β which provide adequate class boundary separation, however constraints must be defined to limit β solutions to favour an optimal boundary location. Traditionally, the optimal separating hyperplane is defined as the hyperplane with maximum distance between said hyperplane and closest opposing class samples. Of course, such simplistic linearly problem spaces within real world applications are rare, with the vast majority being non-linear, noisy and incomplete datasets with resulting overlapping class distributions. Margin for error or slack is generally built into the formulation of the optimisation problem to allow for slight overlap of opposing samples between the decision boundary. The optimisation problem with margin constraints and slack becomes:

$$\max_{\beta,\beta_0,\|\beta\|=1} M \tag{6.3}$$
$$\text{subject to } y_i(x_i^T\beta+\beta_0)\geq M-\xi_i, i=1,\ldots,N$$

where $\xi=\{\xi_1,\ldots,\xi_N\}, \xi_i\geq 0, \sum_{i=1}^N\leq C$, the slack, measures the distance overlap of incorrect samples from the margin. The constant C bounds the total proportional distance allowed by predictions to lie on the wrong side of the margin. With consideration of the constraint $\|\beta\|=1$, the overall constraint in eq. (6.3) can be reformed as:

$$\frac{1}{\|\beta\|}y_i(x_i^T\beta+\beta_0)\geq M-\xi_i \tag{6.4}$$

or

$$y_i(x_i^T\beta+\beta_0)\geq M\|\beta\|-\xi_i. \tag{6.5}$$

Since any positively scaled multiple of β and β_0 will satisfy eq. (6.3), $\|\beta\|$ can be arbitrarily set to $1/M$. As a result, eq. (6.3) with modified constraint eq. (6.5) can be rewritten as:

$$\max_{\beta,\beta_0} \frac{1}{\|\beta\|} \tag{6.6}$$
$$\text{subject to } y_i(x_i^T\beta+\beta_0) \geq 1-\xi_i, i=1,\ldots,N$$

where the maximization of $\|\beta\|^{-1}$ is the equivalent of the minimization of $\|\beta\|^2$ and thus arriving at the minimisation problem of:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i \tag{6.7}$$
$$\text{subject to } \xi_i \geq 0, y_i(x_i^T\beta+\beta_0) \geq 1-\xi_i$$

producing the final form of the maximum margin classifier, generally associated with support vector machine (SVM) model formulation. The subsequent rearrangement into the Lagrangian dual form thus provides an easily solvable convex optimisation problem through quadratic programming solutions.

In order to facilitate the proposed shifting margin concept, a further hyperparameter coefficient, λ_m, is defined to dictate margin size proportion, $C' = \lambda_m C$. Said coefficient allows for the adjustment of margin size proportional to the original maximum margin through the modification of the original objective function constraints in eq. (6.3) to:

$$y_i(x_i^T\beta+\beta_0) \geq \lambda_m C - \xi_i. \tag{6.8}$$

Repetition of the aforementioned derivation will result in the optimisation problem:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i \tag{6.9}$$
$$\text{subject to } \xi_i \geq 0, y_i(x_i^T\beta+\beta_0) \geq \lambda_m - \xi_i$$

Taking the modified optimisation problem, eq. (6.9) into a Lagrangian dual form to produce a solution would simply result in the elimination of the shifted margin coefficient and result in the same maximisation of the decision boundary margin. However, through optimisation via stochastic gradient descent within a neural network (NN) training procedure, the influence of the shifted margin is maintained. As such, the derivation into a loss function applicable with stochastic gradient descent optimisation can be performed. With consideration of $\xi_i \geq 0$

indicating ξ_i must be a positive real value, the constraints defined by eq. (6.9), $y_i(x_i^T\beta+\beta_0) \geq \lambda_m - \xi_i$ can be redefined as a loss function in a similar manner to that of the hinge loss:

$$\xi_i = \max(0, \lambda_m - y_i(x_i^T\beta+\beta_0)) \tag{6.10}$$

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N} \max(0, \lambda_m - y_i(x_i^T\beta+\beta_0)) \tag{6.11}$$

In regards to our loss function being applied to a deep NN architecture, the model prediction, $\lambda_m - y_i(x_i^T\beta+\beta_0)$ can be obfuscated to simply model prediction, $\hat{y}_i$, with β model parameters updated through traditional NN back propagation means. Further obfuscations, in consideration of this being a NN loss function includes $\frac{1}{2}\|\beta\|^2$ simply being traditional L2 regularisation of model weights, whilst C, a constant factor becomes irrelevant in the partial differentiation with respect to β and L. Consequently, the final shifting margin hinge loss can be simplified to:

$$L = \sum_{i=1}^{N} \max(0, \lambda_m - y_i\hat{y}_i) \tag{6.12}$$

where y are timestep class labels with positive and negative class values being: $y^+ = 1, y^- = -1$. Model predictions are defined as $\hat{y}$ and λ_m is the hyperparameter coefficient dictating margin size where $\lambda_m \in \{\mathbb{R} \geq 0\}$.

6.3.3 Critical Diagnosis Point Penalty

As mentioned in section 6.2, the prediction objective is per-timestep binary prediction, $\hat{y}_t$ of sepsis on said timestep. However, emphasis is placed on detection of the specific timestep where sepsis onset occurs. Post-onset, sepsis is continually present within the patient timeline. Consequently, per-timestep binary prediction past point of onset should remain indicating a positive indication of sepsis. As such, the application objective can be alternatively expressed as a pseudo regression task, predicting the point until said critical sepsis-onset point, t_{sepsis}. Such a secondary objective is not reflected well within the required primary per-timestep binary classification approach of the current application, without consideration of this critical sepsis onset point.

Conversion of the prediction objective into a regression based prediction methodology would remain inappropriate for such an application. Being a real-time streamed application monitoring patient vital signs, a regression based prediction of time-period till sepsis onset requires the internal modelling of future patient vital-signs, significantly increasing the challenge of the overall application, especially the further the time till sepsis onset. The binary

classification of per-timestep sepsis onset with a six hour early lead-up, t_{opt} provides a simplified objective whilst maintaining the aspect of early detection. Such a rationalization of said critical sepsis point however, can still be leveraged into the application.

This chapter proposed the critical diagnosis point penalty function as a means to emphasise said critical diagnosis point by introducing a loss penalty based on timestep distance between initial prediction of sepsis-onset to true optimal prediction point, t_{opt}. Early or late predictions produce a linearly increasing absolute error, driving sepsis-onset prediction towards this optimal prediction point. Said penalty, can be based on true positive, false positive, and false negative per-timestep classifications.

Let $\hat{y}_t$ be final class prediction at timestep, t by the considered model, the overall penalty, L_C is driven by the summation of several distinct penalty functions depending on misclassification type via eq. (6.13). True positive predictions, are dictated by eq. (6.14), four continuous piecewise linear equations with parameters dictated by considered time-periods, t_{early}, t_{opt} and t_{late}. Where t_{sepsis} indicates the zero hour clinical diagnosis point of sepsis onset.

False negatives, indicating a late onset prediction, is similarly dictated via eq. (6.15), comprises of two continuous piecewise function, producing a linearly increasing penalty as predicted initial onset extends past the t_{opt} point.

$$L_C = \lambda_C \sum_{t=0}^{T_i} \begin{cases} L_{TP}(t - t_{sepsis}), & \text{if } \hat{y}_t \text{ is TP} \\ L_{FN}(t - t_{sepsis}), & \text{if } \hat{y}_t \text{ is FN} \end{cases} \tag{6.13}$$

$$L_{TP}(t) = \begin{cases} \lambda_{TP} + \lambda_e, & \text{if } t < m_1(\lambda_{TP} + \lambda_e) + b_1 \\ m_1(t) + b_1, & \text{else if } t < t_{opt} \\ m_2(t) + b_2, & \text{else if } t < t_{late} \\ \lambda_{TP}, & \text{otherwise} \end{cases} \tag{6.14}$$

$$L_{FN}(t) = \begin{cases} \lambda_{TP}, & \text{if } t < t_{opt} \\ m_3(t) + b_3, & \text{else if } t < t_{late} \\ 1, & \text{otherwise} \end{cases} \tag{6.15}$$

where

$$m_1 = \frac{-\lambda_{TP}}{(t_{opt} - t_{early})}, \qquad b_1 = -m_1 t_{opt},$$

$$m_2 = \frac{\lambda_{TP}}{(t_{late} - t_{opt})}, \qquad b_2 = -m_2 t_{opt},$$

$$m_3 = \frac{1-\lambda_{TP}}{(t_{late}-t_{opt})}, \qquad b_3 = -m_3 t_{late} + 1$$

Hyperparameter, $\lambda_{TP} \in \{0 \leq \mathbb{R} \leq 1\}$ are used as a weighting coefficient to dictate balance between true positive and false negative predictions, depending upon dataset class balance, whilst hyperparameter λ_C dictates overall contribution of the penalty function towards the loss function. Hyperparameter λ_e is a small coefficient dictating a loss penalty for too early of a positive prediction.

6.3.4 Negative Reversal Penalty

As mentioned previously, a secondary prediction objective of the proposed methodology is the identification of the optimal point, six hours prior to clinical diagnosis of sepsis. Post-optimal point, no physical medical intervention is provided for a septic patient, whilst physical manifestation of patient improvement after intervention requires longer still. As such, any reversion of an originally indicated septic patient by the application, may not occur within our considered patient timeline. As such, a negative reversal penalty is imposed to eliminate such behaviour. Said penalty, produces a linearly increasing loss penalty dependent on distance from the initial positive sepsis prediction, indicating sepsis onset, to any reversion to a negative prediction state post onset indication.

$$L_N = \lambda_N \sum_{t=0}^{T_i} \begin{cases} t'-t, & \text{if } \hat{y}_t = 0 \wedge \exists\ \hat{y}_{t'} = 1 \in \{\forall \hat{y}_{t'} : t' < t\} \\ 0, & \text{otherwise} \end{cases} \tag{6.16}$$

Let $\hat{y}_t$ be final class prediction at timestep, t by the considered model, and $\hat{y}_{t'}$ be all previous timesteps, t' before the considered current timestep, t. Said linear penalty is simply produced by time interval between predicted sepsis onset and negative timestep prediction, $(t'-t)$. Hyperparameter, λ_N dictates overall contribution of the penalty function towards the loss function and serves to balance elimination of highly late negative prediction reversions whilst allowing marginal shifting of the sepsis onset prediction point at each training iteration.

6.4 Experimental Results & Evaluation

6.4.1 Experimental Procedure

Experimental procedure follows closely to that of the Physionet 2019 challenge and are detailed as follows. Prediction objective will be an hourly binary classification of positive sepsis

occurrence at said timestep, with a six hour early positive lead-up to official clinical diagnosis as established within section 6.2. Said prediction objective applies to both the challenge dataset for direct comparison against challenge participants and the MIMIC dataset, for generalisability evaluation whilst maintaining similar experimental procedure to the majority of studies involving MIMIC.

For challenge datasets A and B, set A will form the training and validation set whilst set B will be kept aside for testing evaluation and vice versa to form the main comparative metrics for participant comparisons. Participant results will be taken from the retrospective study by challenge organisers Reyna *et al.* [206]. Finally a 5-fold cross validation procedure will be performed on the combined A and B datasets. In regards to the MIMIC dataset, a 5-fold cross validation across patients will be performed on the overall dataset.

Additionally, Reyna *et al.* proposes a custom 'utility score' metric for challenge evaluation, emphasising timely correct prediction of sepsis onset within a certain time window within the optimal, six hour early, sepsis prediction point. Said metric, shown in fig. 6.2 produces a linearly increasing positive score contribution as true positive prediction points approach t_{optimal}, whilst negative penalties are applied for both too early and too late a positive prediction post-optimal point. Additionally, false positive predictions incur a constant penalty value of =0.05. Subsequently, true negative predictions do not count towards utility score.

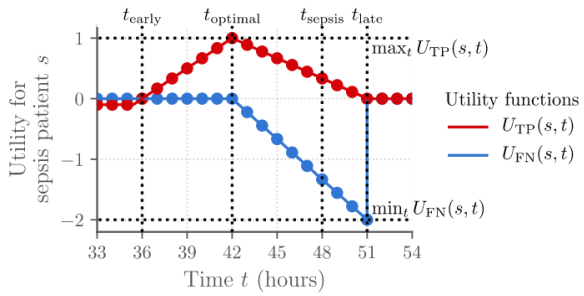

Figure 6.2: Graph indicating the scoring system of the 'utility score' evaluation metric proposed by Reyna *et al.* [206] emphasising closeness of initial indication of sepsis by an application to that of the optimal early clinical diagnosis point. Model predictions are mapped to said utility score function to indicate score gain or loss at each timestep prediction. Of note, a true negative prediction accrues no score whilst a false positive prediction accrues a constant -0.05 utility score.

Utility score collation, bounding and normalisation of an evaluated model is applies as follows:

$$U_{\text{normalised}} = \frac{U_{\text{total}} - U_{\text{inactive}}}{U_{\text{optimal}} - U_{\text{inactive}}} \tag{6.17}$$

As seen, normalisation is based off of a theoretical utility score produced by two model extremes: an optimal classifier with perfect accuracy as the upper bound and an inactive classifier producing fully negative predictions as the lower bound. However, such a normalisation strategy still results in theoretical utility scores of less than zero due to the constant -0.05 false positive penalty. Accordingly, utility score is bounded as, $U_{\text{normalised}} \in \{-0.05 \times N(\text{condition positive}) \leq \mathbb{R} \leq 1\}$.

Model architecture is as indicated within section 6.3.1 following the model formulation of the boosted cascading sub-network training procedure. In regards to training hyperparameters, dropout layers were set at 25% dropout proportion, whilst Adam was used as the gradient optimisation methodology with parameters, $\text{lr} = 0.001$, $\beta_0 = 0.9$, $\beta_1 = 0.999$, $\varepsilon = 1 \times 10^{-7}$ and batch size $= 1000$. Train epochs were set at a maximum of 500 with an early stopping criteria of minimum delta improvement of 0.0001 within the training loss. Hyperparameter selection was performed based off greedy grid search criteria on the initial sub-model and kept the same for each subsequent model. Model selection post training was based on the highest performing epoch weights based off validation loss. Cascade creation stopping criteria was dictated by a similar overall minimum delta model validation loss improvement of 0.0001.

6.4.2 Physionet CinC Challenge Results

Following on are evaluation results for the Physionet dataset as produced by the experimental procedure detailed in section 6.4.1. Full evaluation metrics are provided in table 6.4 whilst relevant metrics with comparisons of the top 5 tanked teams are provided in table 6.5 as reported by Reyna *et al.* [206]. Of note, challenge team results show top scoring metrics of multiple submitted trial runs, whereas the proposed methodology metrics are mean results of 10 overall experimental runs. AUROC and AUPRC results were not reported for the top ranked team within the retrospective study.

As shown, the proposed methodology surpasses all top team results except in set A utility score. Major improvements of up to three times in F1 score and AUPRC can be seen across both test sets. Statistically significant performance improvements can be seen in AUPRC, accuracy and F1 score metrics for both tests sets with other team metrics lying outside of a two standard deviation range, in reference to table 6.4. A traditional two-sample t-test is unavailable due to lack of raw prediction results from each team.

In regards to the fine details, comparisons between state-of-the-art study approaches show the most significant improvements are seen in AUPRC and F1 score metrics with very large

Table 6.4: Overall evaluative metrics of the proposed methodology across the datasets

Metric	Set A	Set B	Set A&B
True Pos. Rate	0.480±0.093	0.533±0.006	0.470±0.105
True Neg. Rate	0.982±0.010	0.985±0.002	0.977±0.019
False Pos. Rate	0.018±0.010	0.015±0.002	0.023±0.019
False Neg. Rate	0.520±0.093	0.467±0.006	0.530±0.105
Pos. Predictive Value	0.374±0.092	0.336±0.038	0.341±0.130
Neg. Predictive Value	0.988±0.002	0.993±0.000	0.990±0.003
False Omission Rate	0.012±0.002	0.007±0.000	0.010±0.003
False Discovery Rate	0.626±0.092	0.664±0.038	0.659±0.130
Accuracy	0.971±0.008	0.979±0.002	0.968±0.017
F1 Score	0.420±0.008	0.412±0.021	0.363±0.058
AUROC	0.855±0.032	0.893±0.026	0.737±0.142
AUPRC	0.391±0.010	0.351±0.042	0.258±0.051

Table 6.5: Physionet CinC 2019 challenge top 5 final leaderboard with proposed methodology results

Team Name	Utility Score		AUROC		AUPRC		Accuracy		F1 Score	
	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B
Proposed Methodology	0.415	**0.450**	**0.855**	**0.893**	**0.391**	**0.351**	**0.971**	**0.979**	**0.420**	**0.412**
Can I get your signature?	**0.433**	0.434	-	-	-	-	0.828	0.888	0.139	0.140
Sepsyd	0.409	0.396	0.811	0.853	0.105	0.119	0.819	0.901	0.131	0.142
Separatrix	0.422	0.395	0.814	0.844	0.102	0.110	0.803	0.882	0.128	0.130
FlyingBubble	0.420	0.401	0.813	0.855	0.108	0.117	0.798	0.878	0.126	0.129
CTL-Team	0.401	0.407	0.806	0.846	0.101	0.116	0.797	0.891	0.122	0.137

Challenge leaderboard is taken from the Physionet challenge retrospective paper by Reyna *et al.* [206]. Of note, several metrics regarding the top performing team, *"Can I get your signature?"* were not provided within the retrospective paper and thus also omitted within this paper's results.

gains in both. Accuracy and AUROC metrics however, show comparatively less improvement with generally equal to marginal improvements in performance results. Combined with large discrpancies between AUROC, AUPRC and F1 Score, all can be equated to the low prevalence rates of sepsis within the dataset. Differences in class weighting parameters on low prevalence datasets emphasising that of high sensitivity over high precision or vice versa produce significant swings in metric values.

The proposed methodology provides superior capability in discerning condition negative class samples due to the proposed novel boosted cascading sub-network architecture, maintaining model sensitivity comparative to current state-of-the-art without the significant impact on model specificity. As such, the proposed methodology provides superior class distinction in significantly imbalanced datasets generally inherent within medical informatics. However, even with a significant improvement as compared to state-of-the-art, there remains considerable margins for improvement with a precision best of 0.374 within the Physionet dataset. With a near 13 to 1 negative to positive class imbalance, such class imbalance still remains an

ever-present non-trivial challenge with potential for future improvement.

6.4.3 MIMIC-III Results

Following on, are the test results of the MIMIC-III dataset showing average test results on a 5-fold cross validation procedure. Each fold iteration comprised of a 3-1-1 fold ratio of training, validation and testing sets respectively. Consistent results from the MIMIC dataset highlight model generalisability towards sepsis detection within varied datasets of significant differing dataset characteristics.

The proposed methodology was comparatively evaluated against several traditional ML methodologies as baselines. As shown in table 6.7, the proposed methodology surpasses all traditional ML methodologies in all evaluation metrics. Of note however, is the marginal non-significant improvements between the proposed methodology and itself without the penalties as described in sections 6.3.3 and 6.3.4. Also of note, is the failure in effective prediction by both deep neural network (DNN) and SVM with F1 Scores of 0.017 and 0.013, indicating significant under-fitting bias towards a negative prediction. Adversely, the aforementioned methodologies present high average accuracy. Such results provide a prime example of the commonly understood failings of accuracy as a sole metric in the evaluation of methodologies.

Additionally, an ablative study on the proposed methodology was performed and evaluated to highlight individual effectiveness and combined strengths of proposed components. Result evaluation of each ablation are detailed in section 6.3, highlights the positive individual contribution of each component. The greatest contribution being the inclusion of the boosted cascading sub-network training procedure producing a 17% improvement towards AUROC score as compared to the baseline LSTM model. Shifting margin hinge loss and the assortment of optimal prediction point penalty functions indicate less of a significant improvement individually, however, provide the best prediction combined together into the proposed methodology.

Table 6.6: Overall evaluative metrics of the proposed methodology on MIMIC-III

Metric	Results
True Pos. Rate	0.706 ± 0.034
True Neg. Rate	0.932 ± 0.003
Pos. Predictive Value	0.275 ± 0.045
Neg. Predictive Value	0.989 ± 0.001
Accuracy	0.925 ± 0.003
F1 Score	0.395 ± 0.052
AUROC	0.787 ± 0.021
AUPRC	0.480 ± 0.042

Table 6.7: Test sesults of various traditional machine learning methodologies on the MIMIC-III dataset

Model	AUROC	AUPRC	Accuracy	F1 Score
Proposed Methodology	**0.787**	**0.480**	**0.925**	**0.395**
LSTM+BCSN+SMHL	0.782	0.436	0.917	0.384
LSTM+BCSN	0.770	0.416	0.902	0.367
LSTM	0.604	0.405	0.859	0.324
Random Forest	0.654	0.283	0.875	0.195
Deep Neural Network	0.494	0.292	0.906	0.017
SVM	0.362	0.237	0.901	0.013
Calvert *et al.* [183]	0.92	-	0.827	0.545*
Nemati *et al.* [213]	0.85	-	0.68	≈0.201*
Desautels *et al.* [14]	0.74	0.28	0.57	0.30
qSOFA[14]	0.77	0.28	0.80	0.39
SIRS[14]	0.61	0.16	0.47	0.24

Results from an ablative study of each component within the study's proposed methodology are shown in the first section. Following are baseline traditional models to highlight overall dataset difficulty in the second section. Model results from literature with similar objectives are provided in the third section of this table. Such results are separate from evaluation due to major differences in experimental procedure and data preparation and thus not fully comparable. QSOFA and SIRS as baseline predictors of non-early sepsis. BCSN: Boosted Cascading Sub-Networks, SMHL: Shifting Margin Hinge Loss. *Calculated using reported secondary statistics

6.5 Conclusion

Within this chapter, we propose and demonstrate a novel DL based methodology as a predictive application for the early detection of sepsis onset for continuously monitored patients in an ICU setting. Using the novel boosted cascading sub-network training procedure, augmented with a shifting margin hinge loss and a collection of optimal prediction point penalty functions, the proposed methodology was able to significantly outperform current state-of-the-art applications within the medical informatics field through the Physionet challenge dataset. The proposed methodology demonstrates effective dataset generalisability through maintaining effective prediction performance into the vastly different dataset characteristics of the MIMIC dataset. Each component within the proposed methodology demonstrates effective individual contributions through ablative study.

Superior performance comes with shortcomings within the proposed methodology. Most significantly, as is apparent with all DL based modelling applications, there comes issues of high model complexity resulting in substantial training times compounded with the training of multiple sub-models and extensive size of medical informatics datasets. With highly disparate pre-processing required between the two medical datasets, Physionet and MIMIC, extensive retraining of models are required to achieve good results. Additionally, such complexity also gives rise to minimal understanding of the 'black box' that is a DL architecture. Such understanding is crucial within the critical field of medical care to ensure validation and trust.

Continued development in ML visualisation and understanding coupled with data-mining and feature ranking provide avenues into significant advancements in smart clinical monitoring and decision support systems.

As touched upon in section 6.4.2, slight differences in data pre-processing pipelines and experimental procedure between similar application studies results in highly divergent evaluation test-sets and resulting performance metrics even when applied within the same underlying medical dataset. The resulting reported standard performance metrics of studies are thus often incomparable to each other. Curated challenge datasets such as Physionet, through highly standardised data preparation and objectives, allow for dependable comparisons between the vast array of application studies available.

The ablative study, presented in table 6.7, highlights the individual effectiveness of each novel component of the proposed methodology. As shown, the significant improvement to predictive performance is provided by the inclusion of boosted cascading sub-networks providing a 0.17 improvement in area under reciever operating characteristic (AUROC). Whilst all individual components provide overall improvements in both sensitivity and specificity, the proposed penalty functions provide the least improvement. Such results are expected due to the design objective being the elimination of a highly small sub-set of misclassifications around the aforementioned 'critical diagnosis point'.

In conclusion, in such a domain as electronic health record (EHR) based decision support system applications; the relative scarcity of highly specific conditions and diseases within a large population set of incoming ICU patients presents highly challenging issues of data bias. This chapter highlights a potential solution through the application of ensemble based ML approaches through the novel boosted cascading sub-network approach. Whilst the presented results indicate great promise, there still remains drawbacks to such a highly positive result. Greater effort is needed within the wide domain of ML based studies applied on EHRs for more consistent approaches to data preparation and result evaluation. Further study evaluation is needed on real-world impact of such applications within an actual ICU environment whilst greater emphasis is required on 'explainable AI': evaluating and understanding the decision making process in conjunction to overall model performance.

Albeit, this chapter highlights there is significant potential for modern statistical data analysis within such complex domains as human healthcare.

Chapter 7

Linear Aggregation Kernel Based Feature Ranking: Identifying Predictive Indicators within Electronic Health Records

Contents

7.1 Introduction

Within this chapter, we return to the application of novel embedded feature selection approaches for the discovery of relevant biomarkers. With such a rich diversity of information available in electronic health records (EHRs), current studies leveraging EHR based predictive modelling feature a surprisingly limited selection of predictors[28]. Many traditional modelling applications approach such large feature dimensionality through limiting complexity to a manageably small fraction of variables dependent on domain and objective, dictated by statistical methodology or prior domain/expert knowledge[28]. Such approaches allow for lighter-weight, manageable and more comprehensible analysis and conclusion; however suffer from a disconnect between model performance and relevant feature selection.

To this end, we seek to incorporate lessons learned in the previous chapter regarding effective application of recurrent based neural networks (NNs) for EHRs within an embedded feature selection context. We present a novel deep learning approach incorporating high-dimensional EHRs to model and predict several clinical applications. To combat prevalent complications of incomprehensibly large feature counts preventing practical clinical application, we also incorporate simultaneous culling of insignificant predictors pertaining to the prediction task at hand. Consequently significantly reducing the vast collection of predictors to only a handful, relevant to the clinical task being considered whilst still maintaining predictive capabilities able to outperform traditional statistical modelling and established clinical diagnosis methods.

Prior studies have surprisingly been highly limited in predictor utilization in comparison to the high-dimensional potential of EHRs. Recent systematic reviews of the literature highlight the median count on variable use at close to only 30 variables[28, 32]. A stark contrast to examples such as the common International Classification of Diseases (ICD) system, codifying and hierarchically classifying 14,000 possible unique medical terms[54].

Further indications highlight limited multi-centre validation across multiple disparate EHR systems in the assessment of multi-objective applications[28, 32] and emphasise such importance in good-practice validation[214, 32].

The open-source Medical Information Mart for Intensive Care dataset (MIMIC) dataset[57], and resulting intensive care unit (ICU) based applications, has exhibited considerable popularity amongst recent literature[23, 43]. Such popularity explained by said records being of a higher degree in detail and frequency, required in an ICU department, compared to a less critical-care based setting[57]. Consequently, MIMIC presents a highly desirable vali-

dation tool via rich patient data with non-trivial clinically significant applications and a large selection of recent, comparable, associated literature.

We present several contributions. We introduce a novel embedded feature selection deep learning approach using sparse regularized linear aggregation kernel layers, employing large high-dimensional EHRs. Said aggregation layers reduce large feature spaces of over 1500 unique features down to 5 linearly condensed embedding coefficients able to predict clinically significant EHR based applications whilst maintaining human comprehensibility.

We validate the effectiveness of the proposed methodology through two disparate EHR datasets applied within previous chapters. Specifically, we validate ICU patient data on the MIMIC dataset to predict severe sepsis at a higher performance metric and at a earlier period than that of current established clinical practice. We secondly validate general practice (GP) patient data on the Secure Anonymised Information Linkage (SAIL) dataset to predict dementia onset prior to officially recorded dementia indications.

Finally, we present potentially significant prediction indicators for each application, based on the resulting feature reduction results. Through said features, we highlight already medically established predictors as heuristic proof of feature selection capabilities and unexplored indicators potentially warranting future novel clinical study. Of interest, biomarkers highlighted in table 7.6 within this chapter regarding dementia follow closely to that of chapter 5 selected features, implying indications of feature selection consistency and relevance.

7.2 Datasets & Outcomes

We evaluate on two previously detailed EHR datasets, each with separate clinical applications. We firstly apply the proposed model to highlight indicators of, and predict, clinical suspicion of severe sepsis within an ICU setting using the MIMIC-III EHR dataset[57] as validated in chapter 6. MIMIC-III is characterised, in our comparative case, as a short-term, high-frequency dataset comprised of vital statistics, lab test results and patient demographics with an emphasis on time-critical diagnosis of severe sepsis. In contrast; SAIL contains coarser collections of diagnostic records, medication, primary & secondary care records, amongst other information presenting long-term, low-frequency, sparse data characteristics.

7.2.1 Sepsis: MIMIC-III

7.2.1.1 Patient Outcome

Patient outcome definitions within the MIMIC dataset follow exactly as that of chapter 6 to ensure consistency. We seek to predict development of sepsis at a 6 hour earlier time-period of than that of recorded clinically indicated onset. Clinical indication being defined in line with Sepsis-3 clinical criteria[170] as either:

1. Recorded two point deterioration in quick SOFA (qSOFA) score.
2. Indications of clinical suspicion as blood culture testing and 72 consecutive hours of IV antibiotic administration within a certain time period as follows:
    a) Prescribing of IV antibiotics followed by blood culture testing within 24 hours.
    b) Blood culture testing followed by prescribing of IV antibiotics within 72 hours.

or the earliest indication if both are present. Said criteria being a commonly implemented sepsis indication in recent machine learning (ML) based EHR modelling literature[32, 206].

7.2.1.2 Dataset Processing

In regards to data pre-processing and representation structure, a unique individual can be associated to multiple hospital admissions or 'spells', each a collection of timestamped patient treatment information with over-arcing demographic information. Duration between admissions can span over years with independent admission reasoning and as such, each patient spell has been treated, in essence, as unique and independent to other patient associated spells to ease processing.

Medical events were binned into hourly time-steps of categorised features, representing the set of all unique medical events. Features containing multiple numeric measurements with an hourly bin were condensed to a median value. In cases of missing measurements, values are linearly interpolated between known measurements, otherwise are set to zero in all other cases. Non-numeric features are recorded as counts of hourly occurrences containing unique timestamps. MIMIC-III exceptionally defines ages over 89 as 300 for privacy concerns and has correspondingly been set as 90 for our application. All features, now subsequently transformed to numeric data types with no missing values, were then normalized. Patient outcome, as defined in section 7.2.1.1, is assigned as a positive or negative binary label sepsis indication at each hourly time-step.

Spell timelines were sliced to only relevant periods prior to initial sepsis indication. Spells containing a positive sepsis indication were cut to include everything prior to 12 hours after initial positive indication. Spells containing no positive sepsis indication were not sliced. Spells of less than 1 hour were removed. Rare medical events were removed in ascending order of frequency to a cumulative 1% of total event occurrences. Of 5852 unique medical events, selection of the top cumulative 99% highest frequency unique events results in a final dataset of 1517 unique events presented as features and 53,811 spells.

7.2.1.3 Population Characteristics

Final pre-processed dataset consists of 53,811 spells. Proportions for outcome and gender are shown in table 7.1 with overall proportion of negative and positive indications of sepsis as 48,295 spells (89.7%) and 5516 spells (10.3%). Population characteristics are summarized in fig. 7.1. Patients with a positive indication of sepsis present comparatively longer lengths of stay with median durations of 64 and 49 hours respectively. Significantly greater IQR, and $Q3$ ranges highlight generally greater possibility of complications requiring continued care for septic patients as compared to non-septic. No significant differences can be seen between patient sex and spell duration.

Of sepsis positive patients, median time from admission to sepsis indication is 41 hours while proportion of positive patients with indications of sepsis immediately upon hospital admission (Hour 0) is 18.85% (1040 spells).

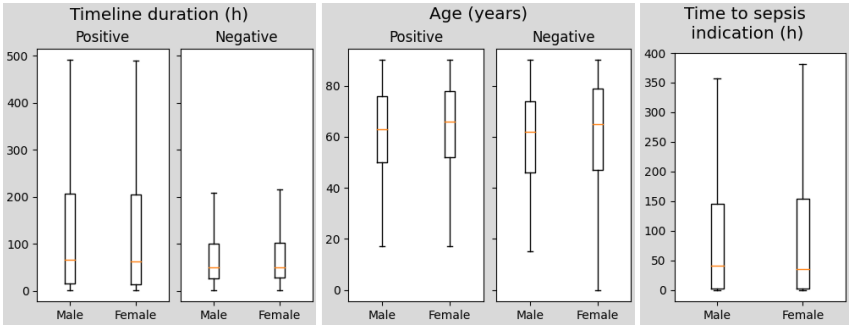

Figure 7.1: Box plot (left) highlighting patient age at admission date, split between categories of sex and positive/negative indication of sepsis during spell. Median ages between categories are 63, 66, 62, and 65 years respectively. Box plot (middle) of overall spell duration of patients from admission to release in hours. Durations are split by categories of sex and positive/negative indication of sepsis during spell. Box plot (right) indicates time in hours from admission to initial indicated suspicion of sepsis. Median time to suspicion is 41 and 36 hours for males and females respectively.

Table 7.1: Sepsis patient proportions for the MIMIC dataset

	Positive	Negative	Total
Male	3167 (0.57)	27056 (0.56)	30223 (0.56)
Female	2349 (0.43)	21239 (0.44)	23588 (0.44)
Total	5516 (0.10)	48295 (0.90)	53811 (1.0)

Male/Female ratios are shown in brackets. As shown, there exists an overall higher proportion of male population in our dataset. There is no significant difference in population sex proportions between patient sepsis categories. A significant imbalance of patient sepsis categories can be seen at only 10% of overall population having sepsis indications.

7.2.2 Dementia: SAIL

7.2.2.1 Patient Outcome

Outcome modelling objective is prediction of dementia diagnosis as recorded on a patient timeline. Class labels are allocated per timestep (weekly) with positive labels allocated starting on the earliest indicated dementia diagnosis date and continuing on to end of patient timeline. Diagnosis of dementia is identified by all NHS read codes hierarchically categorised under codes $[E00..]$, 'Senile and presenile organic psychotic conditions' and $[F110.]$, 'Alzheimer's disease'. The condition negative control cohort is selected randomly from the overall population on a 1:3 population size ratio of positive to negative dementia patients with close to equal patient sex distribution. Post-preprocessing and clean-up of invalid entries reduces said proportion to that shown in table 7.2.

7.2.2.2 Dataset Processing

In terms of SAIL databank datasets of relevance to the task at hand, we utilize the GP dataset and the hospital Patient Episode Database for Wales (PEDW) dataset. Pre-processing of the PEDW dataset remains similar to that of MIMIC (see section 7.2.1.2), consisting of hospital spells, collections of timestamped patient treatment information. The GP dataset contains uncollated timestamped patient treatment information. An individual across both datasets are linked via a unique ALF coding system.

Medical events were binned into weekly time-steps of categorised unique features. Multiple numeric measurements of a feature within a week were condensed to a scalar median value. Missing measurements of said measurement events are linearly interpolated between known measurements, otherwise are set to zero. Non-measurement medical events are recorded as counts of occurrences with unique timestamps within a daily bin to eliminate duplication. Features were subsequently normalised. Patient outcome as defined within the previous section

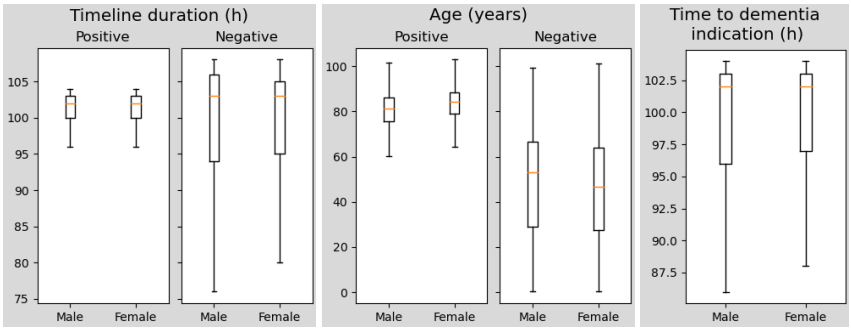

Figure 7.2: Box plot (left) of overall timeline duration of patients in weeks. Durations are split by categories of sex and positive/negative indication of dementia during timeline. Median timeline durations show a slight increase for positive dementia patients compared to negative at 102&102 weeks and 103&103 weeks respectively. Box plot (middle) highlighting patient age at dementia onset, split between categories of sex and positive/negative indication of dementia. Median ages between categories are 81, 84, 53, and 47 years respectively. Box plot (right) indicates time in hours from record start to initial indication of dementia onset. Median time to onset is 41 and 36 weeks for males and females respectively.

is dictated as a positive or negative binary indication of dementia discovery at each weekly time-step.

7.2.2.3 Population Characteristics

Final pre-processed dataset consists of 30,513 patient histories with class proportions shown in table 7.2. Overall class balance of positive and negative dementia patients is considerably more equal than that of the sepsis SAIL dataset at 41% and 59% respectively. Patient timeline characteristics between sex and class are shown in fig. 7.2. As seen, there exists comparatively little differences between patient sex and described characteristics. Unsurprisingly, median age at date of sepsis diagnosis is indicated close to 80 years of age, a stark contrast to a larger distributed general population median of 50 years. Of interest, timeline duration *IQR*, and *Q*2&3 ranges are greatly reduced in the dementia positive population, a stark contrast to the literature highlighting increased hospitalization and institutionalization rates for those suffering from dementia.

7.3 Methodology

Emphasis is placed on producing human-comprehensible predictor rankings whilst maintaining effective modelling capability. Predictor ranking is achieved through novel sparse regularized

Table 7.2: Dementia population proportions for the SAIL dataset

	Positive	Negative	Total
Male	4423 (0.35)	7314 (0.41)	15127 (0.38)
Female	8072 (0.65)	10704 (0.59)	15386 (0.62)
Total	12495 (0.41)	18018 (0.59)	30513 (1.0)

Male/Female ratios are shown in brackets. As shown, there exists an overall higher proportion of female population in our dataset (62% and 38%). There is a larger proportion of positive/negative dementia in females than in males. Overall dementia balance is 12,495 individuals with dementia (41%) and 18,018 without (59%).

weighted linear aggregation kernel layers to achieve large scale linear feature reduction and embedding whilst being easily analysed through traditional statistical analysis. Feature sparsity is achieved through sparse regularization functions penalizing large coefficient weights. Said aggregation kernels are thus attached to a prediction model, which in our case, are a recurrent based NN model to be trained collectively. Post training, the resulting sparse linear embeddings are thus fed into recurrent long short-term memory (LSTM) layers as our deep learning (DL) model, leveraging the longitudinal, time-series component of EHR data to achieve effective predictive performance on our previously established patient outcomes. Proceeding on, we refer to said model as sparse linear feature reduction (SLFR).

7.3.1 Weighted Linear Aggregation Kernels

We introduce the weighted linear aggregation layer in detail. A summary visual diagram is presented in fig. 7.3. The objective of our proposed aggregation layer being significant bottlenecking in information capacity to emphasise efficient data embeddings. With the inclusion of a sparsity inducing weight regularisation loss penalty, feature embeddings are encouraged towards eliminating feature connections by driving weights towards zero. Through the linear combination of features at each kernel, feature importance can be simply back-propagated to individual features.

Applied on a per time-step, t, basis within a spell timeline, we take our $1 \times N$ vector of patient features, x^t, as input for our aggregation kernel layers. All processes are done on a per time-step basis and, for clarity, is assumed on all relevant equations. Consequently, x^t, is our vector of input patient values and y^t our binary patient outcome value at time-step, t, is simplified to x and y respectively. Feature size of the input vector is dictated as N. Each kernel is based on a weighted linear summation of every feature into a singular linearly aggregated

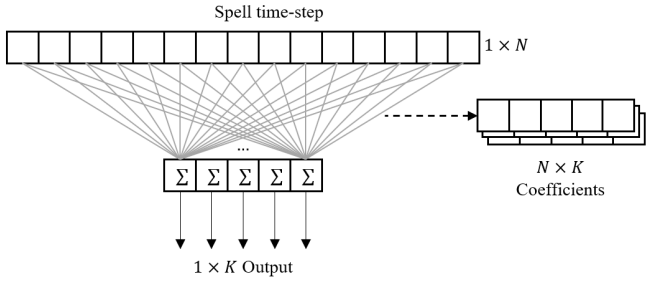

Figure 7.3: Graphical representation of the proposed weighted linear aggregation kernels for linear reduction of feature space. Each kernel produces a linear weighting of each feature before reduction into a significantly smaller embedding space. Combined with sparse regularization and significant information capacity bottlenecking, each kernel produces a unique sparse weighting vector with many features contributing zero weight to the output kernel activation. Model prediction optimization ensures only relevant features are weighted highly and remain.

embedding feature as follows:

$$x'_k = \sum_{i}^{N} w_{i,k} x_i \tag{7.1}$$

resulting in a linear transformation reducing our $N \times 1$ input vector to a scalar output value, x'_k. Each kernel, k, with respective learned coefficients, $w_{i,k}$, one for each input feature, i, linearly weight each unique medical event before reduction through summation to a scalar value. The user-defined K number of kernels produce our new $K \times 1$ output embedding, resulting in a high-dimensional input consisting of potentially thousands of features able to be reduced to K number of embedded features. Consequently, K further controls feature condensation, of which $K = 5$ kernels was heuristically found to effectively balance between model performance and sufficient reduction.

The use of several uniquely initialised kernels allows for unique embeddings across each kernel. Each aggregation kernel can thus be assumed as a unique set of related aggregated medical events with corresponding positive or negative influence towards the remainder of the prediction model.

Through significant bottlenecking in information capacity, combined with model training and optimization for patient outcome prediction, our model is encouraged to produce highly efficient data embeddings able to model subsequent patient outcome. We seek to encourage reduction through elimination of irrelevant features and emphasis of a select set of sparse features through a sparsity regularization function. The resulting sparse linear coefficients allow for simplistic evaluation and visualisation of feature importance.

7.3.2 Sparse Regularization

In combination with the aforementioned aggregation kernel layers, we introduce a regularization technique to encourage weight sparsity through loss penalization of large magnitude kernel coefficients $w_{i,k}$ produced by our proposed aggregation kernels. Feature sparsity is encouraged by driving said coefficients towards zero. Being a linear weighting coefficient, zero valued coefficients effectively filter out irrelevant features within our patient outcome modelling objective.

With the objective of encouraging coefficient sparsity and magnitude minimization, we take each k vector of aggregation kernel coefficients produced by section 7.3.1, $w_{i,k}$, and increasingly penalize each individual coefficient, w_i, magnitude. As such, we propose a normalized inverse exponential function as a loss penalization function on coefficient magnitudes:

$$L(w) = \lambda_1 - \frac{\lambda_1}{N} \sum_{i=0}^{N} e^{-\lambda_2 |w_i|} \tag{7.2}$$

with tuning hyper-parameters $\lambda_1 > 0$ as overall penalty function weight and $\lambda_2 > 0$ as a driving coefficient controlling penalty emphasis on small to large magnitude kernel weight coefficients. As seen, absolute magnitude kernel coefficients, $w_{i,k} \in \mathbb{R}$ are mapped onto a parametrized inverse exponential function, increasing in penalty magnitude as kernel coefficients increase towards a maximum limit, $L(w) \to \lambda_1$, via the normalization component of the proposed function. Our resulting vector of normalized penalty components are reduced to a singular mean penalty value attached to a traditional binary classification loss function for stochastic gradient descent. In our case, the loss function used a traditional hinge loss function optimized by the Adam algorithm.

The application of our parametrized penalty function presents tailored behaviour advantageous to our application of sparsity regularization. In reference to fig. 7.4, intuitively, our penalty function follows traditional weight penalty functions such as L_1 and L_2 via producing an increasing penalty towards increasing weight magnitude with an optimal global zero penalty minimum at zero magnitude. However, at large magnitude coefficient values—our penalty loss, whilst presenting a large value, presents a small gradient approaching zero. Consequently, overly large weight magnitudes produce a low inclination towards being driven towards zero; coefficient optimization is thus largely driven by the prediction error loss. Conversely, small coefficient magnitudes with larger gradient, result in a greater overall trend towards magnitude minimization—counterbalanced by importance towards prediction optimization or driven to zero if irrelevant.

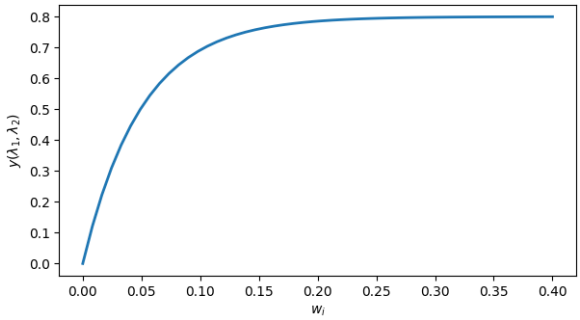

Figure 7.4: Graph detailing the loss penalty curve produced by the sparse regularization function eq. (7.2). The coefficients applied on the proposed methodology are shown as, $\lambda_1 = 0.8$ and $\lambda_2 = 20$. As seen, the regularization function steadily and increasingly penalises weight magnitude. Penalty increase slowly tapers off at extreme weight magnitudes to a set maximum penalty dictated by λ_1. The agressiveness of the curve towards loss maximum is dictated by λ_2 to ensure regularization function spans the entire range of weights within the applied model.

7.3.3 Feature Sparsity

Through application of our aforementioned aggregation kernels with sparsity regularization, several weighting coefficients are produced corresponding to each feature and consequential unique medical event. Analysis of said coefficients allows for the ranking of relevant features and elimination of the vast majority of irrelevant medical events.

As previously mentioned, aggregation kernel amount was set at 5 resulting in 5 corresponding coefficient weightings for each feature. Globally irrelevant features can thus be indicated by a zero absolute sum of corresponding coefficients thus eliminating any influence on the resulting kernel activations as defined by eq. (7.1). Remaining non-zero coefficients, $w_{i,k} \neq 0, w \in \mathbb{R}$ produce a sparse weighted relational matrix indicating importance to individual kernels. Said coefficients can thus be mapped to a weighted many to many relationship between feature and kernel, as seen in fig. 7.5, where each kernel corresponds to a group of related features important to outcome prediction.

7.3.4 Model Architecture

With our proposed linear aggregation layer, longitudinal changes in patient health can be leveraged by an recurrent neural network (RNN) based architecture via recurrent embedded connections through time-steps able to take into account previous data passed into a model. As such, our prediction layer after embedding will consist of LSTM layers. LSTMs are an extension of the RNN unit with respective improvements on long-term memory embeddings[83].

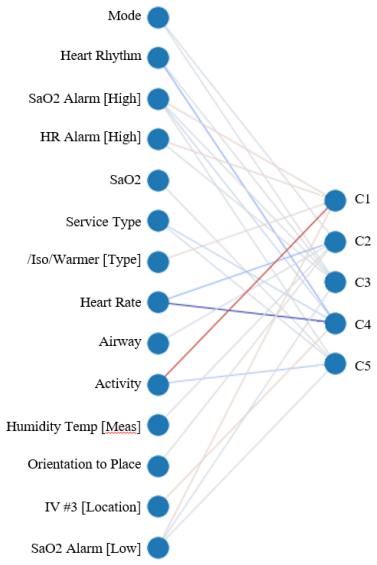

Figure 7.5: Graph representation of an example weighted many to many importance relationship between important features and kernel produced by the proposed methodology. As seen, each kernel produces a group of associated features reduced to a scalar embedding. Sparse regularization ensures only relevant features important in patient outcome prediction. In this case, 5852 total unique medical events of the MIMIC dataset are filtered to only 14 relevant features and linearly reduced to a 5×1 embedding vector.

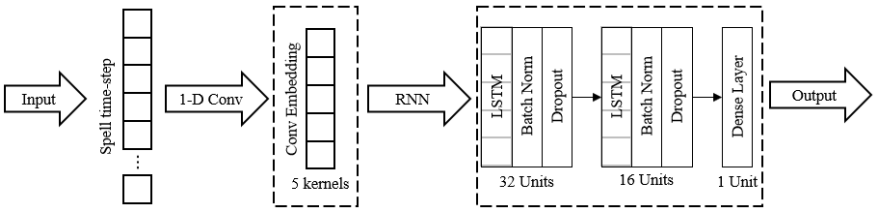

Figure 7.6: Graphical representation of the proposed methodology architecture. Each time-step patient data passes through our sparsity regularized linear aggregation kernel layers for linear compression and sparsification. The aggregation component of our architecture consists of 5 1×1 kernels with no bias or activation component. The resulting embedding is passed through a recurrent architecture consisting of 2 LSTM layers of 32 and 16 units respectively with additional batch norm and dropout regularization. Model output is performed by a singular perceptron dense layer with sigmoid activation for binary classification.

Several regularization techniques were also applied to each recurrent layer, namely: unit dropout, batch normalization and L2 weight regularization. Model optimization was performed using the Adam[215] optimizer on a hinge loss function for binary prediction. Dropout was set at 0.75 ratio throughout. Training was performed using the TensorFlow framework with training hyper-parameters dictated by pattern search optimization. Final model architecture

with relevant hyper-parameters are presented in fig. 7.6. Optimization validation and model testing were achieved using 5-fold cross-validation.

7.4 Results

Following on are testing results for the proposed methodology, SLFR, for both the MIMIC dataset predicting sepsis onset and the SAIL databank dataset predicting dementia onset.

7.4.1 MIMIC-III: Sepsis

Classification performance was evaluated via a 5-fold cross-validation testing approach on the MIMIC dataset. Of which, the dataset was split into 3-1-1 fold ratios as training, validation and testing sets respectively. All results shown are aggregated testing set performance metrics across the 5 fold iterations. Dataset and outcome construction and preprocessing is discussed in section 7.2. Complete results are presented in table 7.3.

Mean area under reciever operating characteristic (AUROC) at sepsis prediction 6 hours prior to clinical indication was 0.787 ± 0.018 with a sensitivity of 0.696 ± 0.029 and specificity of 0.914 ± 0.003. Additional example reported performance metrics of current clinical screening approaches, qSOFA and systematic inflammatory response syndrome (SIRS), are provided for reference[14, 181, 13]. As shown, SLFR exceeds or matches reported AUROC statistics across the 3 related studies. Inconsistent study heterogeneity across dataset preparation, especially in population outcome definitions and inclusion criteria, manifests as comparatively inconsistent performance metrics within studies. Consequently, definitive conclusions are difficult to establish amongst studies.

Table 7.4 and associated heat map fig. 7.7 illustrate coefficient statistics highlighting feature importance within a singular k-fold validation iteration. As discussed in section 7.3, SLFR optimizes 5 kernels to produce sparse coefficient vectors indicating event importance. Insignificant medical events with global coefficient values of 0 were removed from evaluation. Accordingly, of the 1517 unique medical events, 73 events remained significant in sepsis prediction. Kernel coefficient sparsity of the 73 events are portrayed in fig. 7.7 in heat map form. Corresponding event descriptions are provided in table 7.4. Top 20 events are ranked in importance by average absolute coefficient magnitudes across the 5 kernels. Standard deviation of event coefficients across kernels correlate highly with absolute average highlighting low feature importance remaining consistently low across all kernels.

Feature rankings in table 7.4 reveal medical events which follow established medical principles. Events indicative of organ dysfunction: heart, lungs, and consciousness; consistent with the qSOFA assessment procedure are present. Emphasis is placed on heart rate and rhythm with 4 of the top 10 events being relevant, in addition to associated events (3 of top 20) indicating ectopic heart rhythm. Breath sounds, oxygen saturation (Sp02), peak inspiratory pressure (PIP), and head of bed (indicative of ventilator use[216, 217]) associate with lungs. Whilst events, activity and level of consciousness associate with patient consciousness.

Of interest, several established indications of neonatal sepsis are also highlighted. Associations are also seen with gestational age, our top ranked event, and sepsis onset in infants within neonatal intensive care units[218, 219]. Abnormal Moro reflex present within infants is generally indicative of a variety of compromised conditions, including infection[220]. Low birth weight infants (present weight) presents high risk of early onset neonatal sepsis[221, 222].

Table 7.3: Model Performance Statistics for MIMIC: Sepsis

Metric	SLFR Timestep	SLFR Overall	SIRS[14]	qSOFA[14]	SIRS[181]	SIRS[13]	qSOFA[13]
TPR	0.694±0.007	0.696±0.029	0.72	0.56	0.067±0.141	0.464	0.082
TNR	0.988±0.000	0.914±0.003	0.44	0.84	0.740±0.013	0.939	0.996
PPV	0.199±0.003	0.228±0.023	-	-	-	0.133	0.278
NPV	0.999±0.000	0.988±0.000	-	-	-	-	-
Accuracy	0.987±0.000	0.907±0.003	0.47	0.80	-	-	-
F1 Score	0.310±0.004	0.344±0.029	0.24	0.39	-	-	-
AUROC	0.828±0.004	0.787±0.018	0.61	0.77	0.396±0.051	0.79	0.66
AUPRC	0.416±0.003	0.446±0.026	0.16	0.28	-	-	-

7.4.2 SAIL: Dementia

Classification performance was evaluated via a 5-fold cross-validation testing approach on the SAIL dataset. Of which, the dataset was split into 3-1-1 fold ratios as training, validation and testing sets respectively. All results shown are aggregated testing set performance metrics across the 5 fold iterations. Dataset and outcome construction and preprocessing is discussed in section 7.2. Classification performance per timestep produces a sensitivity of 0.694±0.007, specificity of 0.988±0.000 and AUROC of 0.828±0.004 as shown in table 7.5.

Overall patient dementia classification performance is presented in table 7.5 with comparisons against the standard Mini Mental State Exam (MMSE) screening test. As shown, as an automated screening application, SLFR matches, an in some cases exceeds, that of human performed screening approaches. As mentioned previously, with no established gold standard for

Table 7.4: Event Rankings for MIMIC: Sepsis

Index	Abs. Avg.±St.D.	Event
0	0.0578±0.0795	Gestational Age
1	0.0377±0.0754	HR Alarm [High]
2	0.0298±0.0413	Heart Rate
3	0.0280±0.0560	Breath Sounds [L]
4	0.0247±0.0444	Ectopy Type 1
5	0.0196±0.0392	Head of Bed
6	0.0141±0.0171	Ectopy Frequency
7	0.0111±0.0102	Heart Rhythm
8	0.0107±0.0214	Code Status
9	0.0095±0.0132	Heart Rhythm
10	0.0078±0.0079	calprevflg
11	0.0072±0.0143	Moro Reflex
12	0.0063±0.0091	PIP
13	0.0058±0.0096	Present Weight
14	0.0056±0.0068	Sheepskin
15	0.0049±0.0061	SpO2
16	0.0040±0.0079	Activity
17	0.0037±0.0031	GENDER
18	0.0036±0.0061	Ectopy Type
19	0.0035±0.0045	Level of Consciousness

Event rankings are based on feature importance as dictated in section 7.3. Each event is an aggregated overall feature importance metric based on the absolute mean of the 5 feature importance values from each of the 5 kernels. The reported rankings are thus ommitting positive or negative importance weightings.

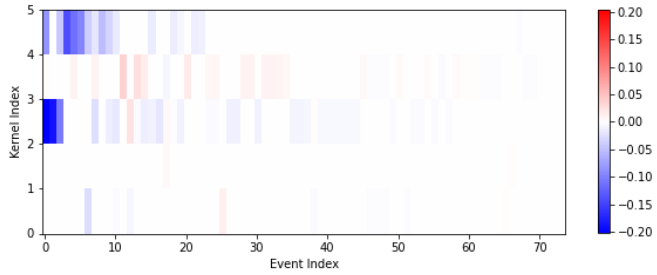

Figure 7.7: Presented is a heatmap of the resulting kernel coefficient weightings produced by the proposed methodology on the MIMIC dataset. Only features with non-zero magnitude coefficients are shown. Each coefficient corresponds to an important relationship between specific unique feature and kernel. Event index definitions are provided in table 7.4. Larger coefficients correspond to greater weighting impact during the linear reduction of each kernel. As seen, there remains significant sparsity within our kernels.

dementia or mild cognitive impairment (MCI) diagnosis, coupled with inconsistent experimental procedure between studies, conclusive comparisons are difficult to establish.

Aggregation kernel linear feature reduction resulted in 43 relevant events out of 2196 total unique events. Top-20 non-zero coefficient values of a single k-fold are presented in table 7.6 with corresponding heat map, fig. 7.8. Of said top-20 features: 3 medical codes reference medical procedures, 1 reference to patient demographics (age), and 16 medication codes. As

seen, top predictive feature, by a significant margin, of dementia indication is patient age. An unsurprising result considering age being a well established risk factor of dementia[223] in addition to being a consistently available patient data-point and having clear correlations with dementia seen in section 7.2.2.3.

In reference to the 16 medication codes, patient medical indications were derived using the British national formulary (BNF)[224] as reference via medication and dosage. The largest collection of associated medications present indications of hypertension, with 4 associated medications including the top 2 medication codes. Correlations between hypertension and dementia prevalence are well known[225, 226].

Further presented medical indications confirmed as risk factors include insomnia[227], renal failure[228], depressive disorder[229], epilepsy[230] and gastrointestinal disorders[231]. Indications of myasthenia gravis are portrayed as an inconclusive risk factor[232]. Indications of ringworm and urinary incontinence associate as common co-morbidities to elderly at-risk or confirmed dementia. No relevant studies were found in a review of literature for correlations between eczema and dementia, presenting a potential avenue of investigation.

Of interest, Haloperidol presents indications of persistent aggression or psychotic symptoms in moderate to severe Alzheimer's dementia and vascular dementia, and consequently, can be inferred as an unforeseen partial ground truth label within our dataset with 110 of 118 patients with prescribed Haloperidol being identified as dementia positive. With consideration however of a larger total of 12,495 positive patients, inclusion of said feature would insignificantly affect performance result validity. The identification of Haloperidol as a significant feature however, presents an interesting unforeseen case, highlighting model capability for effective predictive feature selection.

Table 7.5: Overall Patient Dementia Classification for SAIL: Dementia

Metric	SLFR	Specialist MMSE[233]	MMSE [234]	MMSE [235]
TPR	0.697±0.010	0.761	0.44	0.725
TNR	0.849±0.005	0.886	0.69	0.913
PPV	0.760±0.008	0.893	-	0.769
NPV	0.804±0.006	0.748	-	-
Accuracy	0.787±0.005	-	-	-
F1 Score	0.727±0.007	-	-	-
AUROC	0.762±0.007	-	0.65	0.89
AUPRC	0.761±0.006	-	-	-

Event rankings are based on feature importance as dictated in section 7.3. Each event is an aggregated overall feature importance metric based on the absolute mean of the 5 feature importance values from each of the 5 kernels. The reported rankings are thus ommitting positive or negative importance weightings.

Table 7.6: Event Rankings for SAIL: Dementia

Index	Abs. Avg.±St.D.	Event	Drug Indication
0	0.0104±0.0026	Age	-
1	0.0038±0.0037	Perindopril Erbumine	Hypertension
2	0.0034±0.0029	Doxazosin	Hypertension
3	0.0023±0.0045	Refer To Pathology Department	-
4	0.0021±0.0012	Terbinafine	Ringworm
5	0.0018±0.0018	Temazepam	Insomnia
6	0.0016±0.0013	Haloperidol	Psychotic Symptoms In Dementia
7	0.0016±0.0021	Solifenacin Succinate	Urinary Incontinence
8	0.0015±0.0021	Amitriptyline	Major Depressive Disorder
9	0.0012±0.0015	Adcal-D3 Lemon	Renal Failure
10	0.0010±0.0019	Temazepam	Insomnia
11	0.0009±0.0012	Oilatum Plus	Eczema
12	0.0008±0.0016	Mebeverine Hydrochloride	Gastrointestinal Disorders
13	0.0008±0.0010	Candesartan Cilexetil	Hypertension
14	0.0008±0.0007	Preventive Procedures	-
15	0.0008±0.0015	Co-Codamol	Moderate Pain
16	0.0007±0.0015	Perindopril Erbumine	Hypertension
17	0.0006±0.0013	Prednisolone	Myasthenia Gravis
18	0.0006±0.0011	Epilim Chrono	Epilepsy
19	0.0006±0.0007	Minor Surgery Done - Injection	-

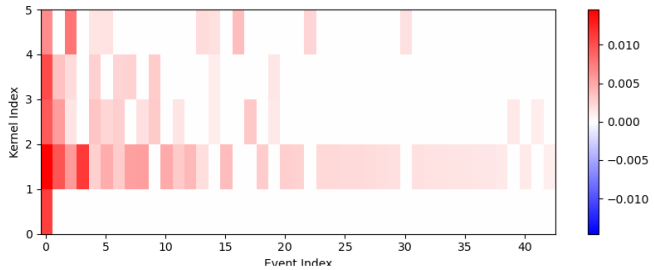

Figure 7.8: Presented is a heatmap of the resulting kernel coefficient weightings produced by the proposed methodology on the SAIL dataset. Only features with non-zero magnitude coefficients are shown. Each coefficient corresponds to an important relationship between specific unique feature and kernel. Event index definitions are provided in table 7.6. Larger coefficients correspond to greater weighting impact during the linear reduction of each kernel. As seen, there remains significant sparsity within our kernels.

7.5 Discussion

We presented a novel automated deep learning approach to simultaneous feature reduction and prediction of multiple patient outcomes in multiple disparate high-dimensional EHR datasets. Specifically, sepsis onset prediction on the MIMIC III ICU patient dataset and dementia onset prediction using the SAIL dataset consisting of GP and hospital patient records. The proposed methodology was able to match or outperform reported performance in literature of traditional, manual, human performed clinical approaches to screening and diagnosis.

Consistent performance is demonstrated on two highly disparate EHR datasets with contrasting patient time-spans and frequency, data characteristics, cohort demographics and modelling objectives highlighting effective model adaptability. Such consistent performance on-par with or exceeding current established clinical procedure, in addition to many other similar EHR based ML application literature[23] highlights the potential positive impact of an automated ML based screening application or clinical decision support system for significant improvements in care quality and patient outcomes[182].

Most significantly, our contributions with the proposed model lie in its capability to perform feature reduction down to small subset of predictive features, highly relevant to the prediction objective at hand. Within the considered EHR datasets, there exists a substantially large set of potential unique medical events within a patient medical timeline, numbering in the thousands, with highly infrequent occurrences. With little to no human intervention on dataset preparation or expert hand-selection of features, the proposed methodology was able to select only a small handful of important predictive medical events proven to have clinical relevance in past literature.

Our study and proposed methodology have several intrinsic and extrinsic limitations. The use of EHRs (with all aforementioned advantages of being longitudinal, patient centric, and with large diverse predictors and sample sizes) present many resulting limitations. Direct comparison and evaluation across studies remains difficult with non-heterogeneous datasets in regards to diverging cohort selection criterion, study design, regional population biases and potential lack of clinical diagnosis gold-standards. The disconnect between EHR patient characteristics and real world population characteristics[236] result in potential invalid, biased associations. Data sparsity and loss to follow-up, in addition to the related and opposing issue of informative observations, present sometimes incomplete and biased observations into a patient's health.

Application and validation on two distinct datasets with differing modelling objectives highlighted good model adaptability. However, true external validation on disassociated EHRs with same prediction objective as opposed to internal EHR cross-validation would emphasise true generalizability and reliability regardless of population and data characteristics. Such validation remains a difficult prospect due to, again, lack of data heterogeneity between EHRs requiring significant manual pre-processing and data association. The use of the MIMIC and SAIL datasets alleviates such issues being each comprised of associated records from multiple primary and secondary care providers however, true external validation (the complete separa-

tion of hospital systems for training and testing) is lacking.

The resulting highlighted predictive event rankings for both sepsis (table 7.4) and dementia (table 7.6) prediction present interesting aspects for discussion. As discussed in section 7.4, almost all highlighted events demonstrate known associations to their predictive objective in question. Interestingly, not highlighted are medical events commonly directly measured in patient diagnosis—such as the established clinical criteria of the qSOFA system in the case of sepsis prediction—however, most highlighted events can be traced via secondary associations to the same approach of organ failure assessment. Similar observations can been seen in dementia prediction with the majority highlighted predictors being prescribed medications as opposed to the direct indicated conditions present within patient histories. Such behaviour follows closely to the proposed methodology presented in chapter 5.

Chapter 8

Conclusions and Future Work

Contents

8.1 Conclusions

This thesis has presented and explored embedded feature selection approaches through the application of temporal-focused recurrent neural network (RNN) based modelling approaches. Such approaches aim to drive neural networks (NNs) away from the opaquely "black box" to be more transparent, presenting methodologies able to push validation through understanding the trained model. Simultaneously, such applications are desirable and of greater feasibility within a health informatics and clinical application mindset, emphasising explainability, simplicity and human capable validation of modelling approaches.

We highlight the dominating challenge of feature redundancy, spurious feature-feature correlations, and overly broad electronic health record (EHR) domains, detailing every medical event present within the individual's health timeline. Such a wide lens on human health simultaneously promises great potential for exploration of novel risk-factors via data mining approaches whilst also presenting significant challenges in effective, feasible, and comprehensible analysis of such big data; as highlighted within this thesis.

The utilization of deep learning (DL) and temporal based machine learning (ML) modelling approaches show great promise in modelling capability of EHRs. Of particular emphasis is ensemble based approaches such as the proposed snapshot ensemble in chapter 6 and boosted cascade approach in chapter 6; able to produce state-of-the-art predictive capability within current literature. Big data EHRs are a difficult prospect through complex challenges of high-dimensionality, high data sparsity, and significant class imbalances; of which, we present ensemble based novel approaches as a potential solution domain as proven via the case studies within this thesis.

DL remains a highly popular and effective approach to ML based applications within a wide variety of research domains including health informatics. The capability for the processing of self-optimised feature representations able to model and predict challenging non-linear data correlations is a well known concept[5]. We present novel approaches towards leveraging feature representations as effective criteria for embedded feature selection approaches. Chapters 5 and 7 makes use of such concepts; producing highly capable feature selection methodologies able to reduce a significantly large domain of thousands of unique medical events down to tens of relevant features able to maintain predictive capability to an effective degree.

We presented resulting feature relevance rankings and selections highlighted within this thesis. As indicated within respective discussions in relevant chapters, there exists great clinical relevance within the presented feature importance results. Of interest is feature similarities

between case studies of alternative methodologies; in particular Tables 5.4 and 7.6 present features of hypertension as of great importance towards dementia development and risk of hospitalization. Such correlations between hypertension and dementia are well known and studied[225, 226].

8.2 Contributions

The main contributions of this thesis can be seen as follows:

- A comprehensive review of EHR based machine learning applications within the current state-of-the-art literature in the context of dementia risk and of sepsis development. Through which we present current relevance of said topics within the greater medical domain, the challenges of applying ML based approaches to such a clinical objective, the current state-of-the-art methodologies and studies within said domain and finally future research pathways and opportunities for further study, of which we approach in the following contributions.

- A novel approach to feature selection within a use case study of identifying at risk individuals with dementia in encountering a hospitalization event. The proposed methodology consists of an ensemble architecture of deep neural networks (DNNs) trained using the "snapshot ensemble" approach to aid in reducing over-fitting and perturb feature weighting in combination with novel entropy weight regularisation to produce sparse feature representations. Such representations are thus applied as ranking and selection criteria to produce a final selected feature-set.

- A novel architecture and training approach towards alleviating issues of high-dimensionality, data sparsity, and class imbalance to produce a prediction tool for sepsis development able to outperform the current state-of-the-art approaches within literature according to the PhysioNet 2019 CinC Challenge. The proposed approach consists of a continuation of the ensemble approach detailed previously, incorporating a novel boosted cascading architecture approach.

- A novel embedded feature selection methodology incorporating 1D convolutional kernels in combination with long short-term memory (LSTM) recurrent networks and a novel weight sparsity regularization approach to produce sparse feature representations within the convolution kernel. Through which, we extrapolate feature importance, or

lack-thereof, to produce a final set of highly relevant features. The proposed methodology was validated through case study using both clinical objectives of predicting sepsis development and separately, dementia development across two uniquely characterised EHRs.

- We present a complete list of discovered biomarkers found to be of relevance towards the proposed case studies. Through which, we analyse the clinical relevance and novelty value to highlight potential avenues of novel clinical research and validation. Said list highlights a significant proportion of known risk factors relevant towards our case studies, highlighting model selection capability whilst several novel medical events were highlighted with minor to no studies exploring such correlation.

8.3 Future Work

The adaption of DL based temporal models applied with embedded feature selection approaches which utilize EHRs for modelling of patient outcomes has a large scope; the wide degree of health information available within EHRs promise both great potential and great challenge. Further application domains within health informatics such as rehospitalisation risk analysis, patient risk and length of stay prediction, and the wide sub-domain of disease diagnosis are ideal for risk-factor and biomarker discovery and automation. With the unprecedented development of SARS-CoV-2 (COVID-19) into a global pandemic, affecting every facet of life, risk-factor analysis and modelling of newly discovered disease trajectories through such proposed methodologies are of significant interest and popularity within domains of research and even of the public eye. With current predictions of continued acceleration of information technology capabilities in all aspects of life, such data-mining approaches extend past that of a purely health informatics domain. Such vast quantities of human data require significant automation and reduction to produce value out of such data.

The development and application of the highlighted field within this thesis remains relatively young. The DNN remains a 'black box' in terms of understanding and validation of such models whilst also presenting state-of-the-art predictive capabilities. There remains much work to be done to improve such understanding required for general acceptance within the evidence based medical approaches for influencing current medical procedure. Such hesitation is, of course, understandable in such a potentially high consequence domain of human medicine.

As highlighted previously, ensemble based approaches towards EHR modelling showed great promise, able to produce state-of-the-art performance capabilities. Said study was limited in scope to purely a predictive clinical objective with no consideration for feature selection and data-mining. The utilization of "snapshot ensembles" further pairs well with feature selection approaches as indicated in chapter 5. The merger of the proposed ensemble cascades with 1D convolution feature selection is an obvious future application.

There exists a significant limitation of the proposed methodologies within chapters 5 and 7 due to the multi-layer and multi-network architectures proposed within each chapter. Feature importance metrics reported remain positive scalar magnitudes indicating relevance and correlation with no indication of positive or negative association. Such indications are possible and available to analyse; however, no assumption can be made on feature representations produced past the initial feature selection layer retaining such associations. As such, positive/negative correlations could potentially vary to an unknown degree during prediction inference. Of interest, the associations produced through evaluation do heuristically match clinical knowledge, but such a indication presents no guarantee.

The current scope of feature selection approaches within this thesis has predominantly focused on global models of feature selection; that is, the evaluation of feature importance across the complete population set to produce a 'global' indication of risk-factors. The highly novel field of local, dynamic feature selection presents great potential for future application. Greater emphasis is placed upon the concept of individualised medicine with feature importance ranking performed on a per patient scale. Such applications have already been proposed and have shown effective results through attention based transformer networks within other domains[237, 238, 239]. Such local feature selection models can further transition towards global feature selection through simple analysis. Of significant interest is local feature selection approaches on the temporal dimension in addition to a per sample basis. Dynamic temporal feature selection indicates feature importance on a time-step level, highlighting longitudinally dynamic feature importance. Such approaches have been presented within the domain of EHRs by Yoon *et al.* [240] following novel similar concepts of attention based networks.

Case study validation across all studies presented within this thesis would further benefit from greater, more diverse applications across a larger variety of population datasets with unique population characteristics. The emphasis of this thesis has been on the Medical Information Mart for Intensive Care dataset (MIMIC) III dataset and the Secure Anonymised Information Linkage (SAIL) datasets with various other similar smaller-scale EHR records

used. Case study validation within chapter 7 presents the ideal case of unique case study and dataset validation whilst studies remain limited to multiple unique datasets within the same case study clinical objective. Model generalisability validation is highly relevant within the health informatics domain due to such potentially diverse health pictures across even national regions, whilst internationally, population characteristics such as education, healthcare, and wealth & deprivation produce highly unique endogeneities. Open access to wide ranges of health information, whilst greatly improved in the recent decade, remains limited due to highly valid concerns of data privacy.

8.4 Closing Remarks

In conclusion, as the age of digital information continually progresses, health information technologies too will advance. The continued digitisation of the individual's health allows for a unprecedented broad-scale view of highly detailed and individualised information across a wide scale of health focuses. EHRs provide substantial opportunity for large-scale, exploratory analysis via modern information technology approaches in the hopes of increasing understanding and improving individualized patient care outcomes and care utilization[21, 22]. Such opportunities however, remain obstructed by significant challenges of big data analysis requiring significant research into tailored approaches able to deal with said challenges. For the data scientist, big data presents significant and unique computational and statistical challenges. For modern society, big data heralds new levels of multi-disciplinary scientific discovery and economic value, promising discovery and analysis of large scale population trends and heterogeneities never before possible with small-scale data.

Bibliography

[1] Cisco Newsroom, *Press Release: Global Internet Traffic Projected to Quadruple by 2015*, 2011. [Online]. Available: `https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2011/m06/global-internet-traffic-projected-to-quadruple-by-2015.html` (visited on 03/30/2022).

[2] H. U. Buhl, M. Röglinger, F. Moser, and J. Heidemann, "Big Data," *Business & Information Systems Engineering*, vol. 5, no. 2, pp. 65–69, 2013.

[3] V. Marx, "The big challenges of big data," *Nature*, vol. 498, no. 7453, pp. 255–260, 2013.

[4] J. Fan, F. Han, and H. Liu, "Challenges of Big Data analysis," *National Science Review*, vol. 1, no. 2, pp. 293–314, 2014.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[6] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[7] R. High, "The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works," *International Business Machines Corporation*, vol. 1, no. 1, pp. 1–14, 2012.

[8] E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research," *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, pp. 48–57, 2014.

[9] A. McCoy and R. Das, "Reducing patient mortality, length of stay and readmissions through machine learning-based sepsis prediction in the emergency department, intensive care unit and hospital floor units," *BMJ Open Quality*, vol. 6, no. 2, 2017.

[10] A. Payan and G. Montana, "Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 2015.

[11] H.-I. Suk and D. Shen, "Deep Learning-Based Feature Representation for AD/MCI Classification," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, vol. 8150, 2013, pp. 583–590.

[12] R. Chen and E. H. Herskovits, "Machine-learning techniques for building a diagnostic model for very mild dementia," *NeuroImage*, vol. 52, no. 1, pp. 234–244, 2010.

[13] R. J. Delahanty, J. Alvarez, L. M. Flynn, R. L. Sherwin, and S. S. Jones, "Development and Evaluation of a Machine Learning Model for the Early Identification of Patients at Risk for Sepsis," *Annals of Emergency Medicine*, vol. 73, no. 4, pp. 334–344, 2019.

[14] T. Desautels *et al.*, "Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach," *JMIR Medical Informatics*, vol. 4, no. 3, p. 28, 2016.

[15] Q. Mao *et al.*, "Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU," *BMJ Open*, vol. 8, no. 1, 2018.

[16] S. Szymczak *et al.*, "Machine learning in genome-wide association studies," *Genetic Epidemiology*, vol. 33, no. S1, S51–S57, 2009.

[17] K. Y. Yip, C. Cheng, and M. Gerstein, "Machine learning and genome annotation: a match meant to be?" *Genome Biology*, vol. 14, no. 5, p. 205, 2013.

[18] Y. Liu *et al.*, "Machine learning in materials genome initiative: A review," *Journal of Materials Science & Technology*, vol. 57, pp. 113–122, 2020.

[19] K. Harron *et al.*, "Challenges in administrative data linkage for research," *Big Data & Society*, vol. 4, no. 2, 2017.

[20] M. A. Bohensky *et al.*, "Data Linkage: A powerful research tool with potential problems," *BMC Health Services Research*, vol. 10, no. 1, p. 346, 2010.

[21] M. R. Cowie *et al.*, "Electronic health records to facilitate clinical research," *Clinical Research in Cardiology*, vol. 106, no. 1, pp. 1–9, 2017.

[22] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, "Mining Electronic Health Records (EHRs)," *ACM Computing Surveys*, vol. 50, no. 6, C. K. Reddy and C. C. Aggarwal, Eds., pp. 1–40, 2018.

[23] A. Rajkomar *et al.*, "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[24] L. Murray *et al.*, "Does prediction of outcome alter patient management?" *The Lancet*, vol. 341, no. 8859, pp. 1487–1491, 1993.

[25] G. C. M. Siontis, I. Tzoulaki, K. C. Siontis, and J. P. A. Ioannidis, "Comparisons of established risk prediction models for cardiovascular disease: systematic review," *British Medical Journal*, vol. 344, no. 1, pp. 3318–3318, 2012.

[26] V. L. Plano Clark, N. Anderson, J. A. Wertz, Y. Zhou, K. Schumacher, and C. Miaskowski, "Conceptualizing Longitudinal Mixed Methods Designs," *Journal of Mixed Methods Research*, vol. 9, no. 4, pp. 297–319, 2015.

[27] S. L. Zeger, R. Irizarry, and R. D. Peng, "On Time Series Analysis of Public Health and Biomedical Data," *Annual Review of Public Health*, vol. 27, no. 1, pp. 57–79, 2006.

[28] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis, "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review," *Journal of the American Medical Informatics Association*, vol. 24, no. 1, pp. 198–208, 2017.

[29] R. a. Lyons *et al.*, "The SAIL databank: linking multiple health and social care datasets," *BMC Medical Informatics and Decision Making*, vol. 9, no. 1, p. 3, 2009.

[30] P. Coorevits *et al.*, "Electronic health records: new opportunities for clinical research," *Journal of Internal Medicine*, vol. 274, no. 6, pp. 547–560, 2013.

[31] F. Davidoff, B. Haynes, D. Sackett, and R. Smith, "Evidence based medicine," *British Medical Journal*, vol. 310, no. 6987, pp. 1085–1086, 1995.

[32] C. Luz, M. Vollmer, J. Decruyenaere, M. Nijsten, C. Glasner, and B. Sinha, "Machine learning in infection management using routine electronic health records: tools, techniques, and reporting of future technologies," *Clinical Microbiology and Infection*, vol. 26, no. 10, pp. 1291–1299, 2020.

[33] C. Shivade *et al.*, "A review of approaches to identifying patient phenotype cohorts using electronic health records," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 221–230, 2014.

[34] C. Bycroft *et al.*, "The UK Biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.

[35] W. Oh *et al.*, "Type 2 Diabetes Mellitus Trajectories and Associated Risks," *Big Data*, vol. 4, no. 1, pp. 25–30, 2016.

[36] R. Horne, J. I. Bell, J. R. Montgomery, M. O. Ravn, and J. E. Tooke, "A new social contract for medical innovation," *The Lancet*, vol. 385, no. 9974, pp. 1153–1154, 2015.

[37] J. Tooke, J. Lundgren, R. Trembath, and J. Iredale, "Stratified, personalised or P4 medicine: a new direction for placing the patient at the centre of healthcare and health education," *The Academy of Medical Sciences*, no. May, p. 37, 2015.

[38] M. E. Porter, S. Larsson, and T. H. Lee, "Standardizing Patient Outcomes Measurement," *New England Journal of Medicine*, vol. 374, no. 6, pp. 504–506, 2016.

[39] K. G. M. Moons, P. Royston, Y. Vergouwe, D. E. Grobbee, and D. G. Altman, "Prognosis and prognostic research: what, why, and how?" *British Medical Journal*, vol. 338, no. 1, pp. 375–375, 2009.

[40] B. A. Goldstein, A. M. Navar, and R. E. Carter, "Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges," *European Heart Journal*, vol. 38, no. 23, p. 302, 2016.

[41] E. Y. H. Tang *et al.*, "Current Developments in Dementia Risk Prediction Modelling: An Updated Systematic Review," *PLOS ONE*, vol. 10, no. 9, G. Forloni, Ed., 2015.

[42] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[43] G. Tsang, X. Xie, and S. M. Zhou, "Harnessing the Power of Machine Learning in Dementia Informatics Research: Issues, Opportunities and Challenges," *IEEE Reviews in Biomedical Engineering*, vol. 13, pp. 113–129, 2019.

[44] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[45] A. Storey, "Living longer: how our population is changing and why it matters," *Office for National Statistics: London, UK*, 2018.

[46] A. L. Huntley, R. Johnson, S. Purdy, J. M. Valderas, and C. Salisbury, "Measures of Multimorbidity and Morbidity Burden for Use in Primary Care and Community Settings: A Systematic Review and Guide," *The Annals of Family Medicine*, vol. 10, no. 2, pp. 134–141, 2012.

[47] T. Lehnert *et al.*, "Review: Health Care Utilization and Costs of Elderly Persons With Multiple Chronic Conditions," *Medical Care Research and Review*, vol. 68, no. 4, pp. 387–420, 2011.

[48] J. Wise, "Polypharmacy: a necessary evil," *British Medical Journal*, vol. 347, no. nov28 1, pp. 7033–7033, 2013.

[49] C. Salisbury, "Multimorbidity: Redesigning health care for people who use it," *The Lancet*, vol. 380, no. 9836, pp. 7–9, 2012.

[50] Medicines and Healthcare products Regulatory Agency, *About Yellow Card*, 2022. [Online]. Available: `https://yellowcard.mhra.gov.uk/the-yellow-card-scheme/` (visited on 01/10/2022).

[51] U.S. Food & Drug Administration, *Questions and Answers on FDA's Adverse Event Reporting System (FAERS)*, 2018. [Online]. Available: `https://www.fda.gov/drugs/surveillance/questions-and-answers-fdas-adverse-event-reporting-system-faers` (visited on 01/10/2022).

[52] K. Kreimeyer *et al.*, "Feature engineering and machine learning for causality assessment in pharmacovigilance: Lessons learned from application to the FDA Adverse Event Reporting System," *Computers in Biology and Medicine*, vol. 135, no. June, p. 104 517, 2021.

[53] A. Mohsen, L. P. Tripathi, and K. Mizuguchi, "Deep Learning Prediction of Adverse Drug Reactions in Drug Discovery Using Open TG–GATEs and FAERS Databases," *Frontiers in Drug Discovery*, vol. 1, 2021.

[54] W. H. Organization, *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. World Health Organization, 1992.

[55] M. Langarizadeh, A. Orooji, and A. Sheikhtaheri, "Effectiveness of anonymization methods in preserving patients' privacy: A systematic literature review," *Studies in Health Technology and Informatics*, vol. 248, no. 6, pp. 80–87, 2018.

[56] R. a. Lyons *et al.*, "The SAIL databank: linking multiple health and social care datasets," *BMC Medical Informatics and Decision Making*, vol. 9, no. 1, p. 3, 2009.

[57] A. E. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, no. 1, 2016.

[58] G. Hripcsak, C. Knirsch, L. Zhou, A. Wilcox, and G. Melton, "Bias Associated with Mining Electronic Health Records," *Journal of Biomedical Discovery and Collaboration*, vol. 6, pp. 48–52, 2011.

[59] C. Lindmeier, *WHO releases new International Classification of DISEASES (ICD 11)*, 2018. [Online]. Available: `https://www.who.int/news/item/18-06-2018-who-releases-new-international-classification-of-diseases-(icd-11)` (visited on 01/17/2022).

[60] B. Blackwell, "Patient compliance," *New England Journal of Medicine*, vol. 289, no. 5, pp. 249–252, 1973.

[61] M. H. van der Wal, T. Jaarsma, and D. J. van Veldhuisen, "Non-compliance in patients with heart failure; how can we manage it?" *European Journal of Heart Failure*, vol. 7, no. 1, pp. 5–17, 2005.

[62] K. C. Rasekhschaffe and R. C. Jones, "Machine Learning for Stock Selection," *Financial Analysts Journal*, vol. 75, no. 3, pp. 70–88, 2019.

[63] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions," *Artificial Intelligence Review*, vol. 53, no. 4, pp. 3007–3057, 2020.

[64] E. Pellegrini *et al.*, "Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: A systematic review," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 10, no. 1, pp. 519–535, 2018.

[65] T. G. Dietterich and E. B. Kong, "Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms," Oregon State University, Tech. Rep., 1995, pp. 0–13.

[66] F. Donald and R. Glauber, "Multicollinearity in Regression Analysis: The Problem Revisited," *The Review of Economic and Statistics*, vol. 49, no. 1, pp. 92–107, 1967.

[67] M. Stephen, *Machine Learning An Algorithmic Perspective Second Edition*. 2014, p. 457.

[68] J. Manyika *et al.*, *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, 2011.

[69] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[70] Tin Kam Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, IEEE Comput. Soc. Press, 1995, pp. 278–282.

[71] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Springer series in statistics New York, NY, USA: 2001, vol. 1.

[72] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[73] A. So, D. Hooshyar, K. W. Park, and H. S. Lim, "Early diagnosis of dementia from clinical data by machine learning techniques," *Applied Sciences*, vol. 7, no. 7, p. 651, 2017.

[74] S.-M. Zhou, H.-X. Li, and L. Xu, "A variational approach to intensity approximation for remote sensing images using dynamic neural network," *Expert Systems*, vol. 20, no. 4, pp. 163–170, 2003.

[75] T. Ching, X. Zhu, and L. X. Garmire, "Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data," *PLOS Computational Biology*, vol. 14, no. 4, pp. 1–18, 2018.

[76] S. M. Zhou and L. D. Xu, "A new type of recurrent fuzzy neural network for modeling dynamic systems," *Knowledge-Based Systems*, vol. 14, no. 5, pp. 243–251, 2001.

[77] D. A. Elizondo, S.-M. Zhou, and C. Chrysostomou, "Light source detection for digital images in noisy scenes: A neural network approach," *Neural Computing and Applications*, vol. 28, no. 5, pp. 899–909, 2017.

[78] S. M. Zhou, "Combining dynamic neural networks and image sequences in a dynamic model for complex industrial production processes," *Expert Systems with Applications*, vol. 16, no. 1, pp. 13–19, 1999.

[79] D. R. Kelley, Y. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, and J. Snoek, "Sequential regulatory activity prediction across chromosomes with convolutional neural networks.," *Genome Research*, 2018.

[80] S. Mc Loone and G. Irwin, "Improving neural network training solutions using regularisation," *Neurocomputing*, vol. 37, no. 1-4, pp. 71–90, 2001.

[81] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object Recognition with Gradient-Based Learning," in *Shape, Contour and Grouping in Computer Vision*, December, vol. 1999, 1999, pp. 319–345.

[82] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[83] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[84] F. Scarselli, M. Gori, Ah Chung Tsoi, M. Hagenbuchner, and G. Monfardini, "The Graph Neural Network Model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[85] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," in *International Conference on Advances in Computing, Communications and Informatics*, vol. 1, IEEE, 2017, pp. 1643–1647.

[86] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, New York, NY, USA: ACM, 2016, pp. 445–450.

[87] A. Fialho, F. Cismondi, S. Vieira, S. Reti, J. Sousa, and S. Finkelstein, "Data mining using clinical physiology at discharge to predict ICU readmissions," *Expert Systems with Applications*, vol. 39, no. 18, pp. 158–165, 2012.

[88] M. Herland, T. M. Khoshgoftaar, and R. Wald, "A review of data mining using big data in health informatics," *Journal Of Big Data*, vol. 1, no. 1, p. 2, 2014.

[89] J. S. Mathias, A. Agrawal, J. Feinglass, A. J. Cooper, D. W. Baker, and A. Choudhary, "Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 118–124, 2013.

[90] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[91] K. Kira and L. A. Rendell, "A Practical Approach to Feature Selection," in *Machine Learning Proceedings*, vol. 256, Elsevier, 1992, pp. 249–256.

[92] Hanchuan Peng, Fuhui Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[93] N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter Methods for Feature Selection – A Comparative Study," in *Intelligent Data Engineering and Automated Learning*, December, vol. 4881, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 178–187.

[94] T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff, "Embedded Methods," in *Feature Extraction*, 9, vol. 60, Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 137–165.

[95] M. Merriman, *A List of Writings Relating to the Method of Least Squares: With Historical and Critical Notes*. Academy, 1877, vol. 4.

[96] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. 2009, vol. 77, pp. 482–482.

[97] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 440–444, 2004.

[98] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.

[99] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[100] J. R. Quinlan, *C4.5: programs for machine learning*. Elsevier, 2014.

[101] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, 2002.

[102] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Fournal on Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[103] Alzheimer's Society, "Dementia UK Report," Tech. Rep., 2018. [Online]. Available: https://www.alzheimers.org.uk/about-us/policy-and-influencing/dementia-uk-report.

[104] S. Banerjee, "The Macroeconomics of Dementia—Will the World Economy Get Alzheimer's Disease?" *Archives of Medical Research*, vol. 43, no. 8, pp. 705–709, 2012.

[105] M. J. Prince, A. Wimo, M. M. Guerchet, G. C. Ali, Y.-T. Wu, and M. Prina, "World Alzheimer Report 2015-The Global Impact of Dementia: An analysis of prevalence, incidence, cost and trends," 2015.

[106] P. D. Sloane *et al.*, "The Public Health Impact of Alzheimer's Disease, 2000–2050: Potential Implication of Treatment Advances," *Annual Review of Public Health*, vol. 23, no. 1, pp. 213–231, 2002.

[107] M. Boustani *et al.*, "Implementing a screening and diagnosis program for dementia in primary care," *Journal of General Internal Medicine*, vol. 20, no. 7, pp. 572–577, 2005.

[108] L. Glodzik *et al.*, "Alzheimer's disease markers, hypertension, and gray matter damage in normal elderly," *Neurobiology of Aging*, vol. 33, no. 7, pp. 1215–1227, 2012.

[109] J. P. W. Bynum, P. V. Rabins, W. Weller, M. Niefeld, G. F. Anderson, and A. W. Wu, "The relationship between a dementia diagnosis, chronic illness, medicare expenditures, and hospital use," *Research and Practice in Alzheimer's Disease*, vol. 10, pp. 160–164, 2005.

[110] K. López-de-Ipiña *et al.*, "On the Selection of Non-Invasive Methods Based on Speech Analysis Oriented to Automatic Alzheimer Disease Diagnosis," *Sensors*, vol. 13, no. 12, pp. 6730–6745, 2013.

[111] J. Escudero, E. Ifeachor, J. P. Zajicek, C. Green, J. Shearer, and S. Pearson, "Machine learning-based method for personalized and cost-effective detection of alzheimer's disease," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 1, pp. 164–168, 2013.

[112] G. M. McKhann *et al.*, "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's & Dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[113] G. López, L. Quesada, and L. A. Guerrero, "Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of speech-based natural user interfaces," in *International Conference on Applied Human Factors and Ergonomics*, Springer, 2017, pp. 241–250.

[114] G. Orrù, W. Pettersson-Yeo, A. F. Marquand, G. Sartori, and A. Mechelli, "Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review," *Neuroscience and Biobehavioral Reviews*, vol. 36, no. 4, pp. 1140–1152, 2012.

[115] L. Mosconi, M. Brys, L. Glodzik-Sobanska, S. De Santi, H. Rusinek, and M. J. de Leon, "Early detection of Alzheimer's disease using neuroimaging," *Experimental Gerontology*, vol. 42, no. 1-2, pp. 129–138, 2007.

[116] T. Ching *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, no. 141, p. 20 170 387, 2018.

[117] G. M. McKhann *et al.*, "The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease," *Alzheimer's and Dementia*, vol. 7, no. 3, pp. 263–269, 2011.

[118] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders*, 5th. American Psychiatric Association, 2013.

[119] M. Lamar, S. M. Resnick, and A. B. Zonderman, "Longitudinal changes in verbal memory in older adults," *Neurology*, vol. 60, no. 1, pp. 82–86, 2003.

[120] T. N. Tombaugh and N. J. McIntyre, "The mini-mental state examination: a comprehensive review.," *Journal of the American Geriatrics Society*, vol. 40, no. 9, pp. 922–935, 1992.

[121] V. G. Valcour, K. H. Masaki, J. D. Curb, and P. L. Blanchette, "The detection of dementia in the primary care setting.," *Archives of Internal Medicine*, vol. 160, no. 19, pp. 2964–2968, 2000.

[122] C. M. Callahan, H. C. Hendrie, and W. M. Tierney, "Documentation and evaluation of cognitive impairment in elderly primary care patients.," *Annals of Internal Medicine*, vol. 122, no. 6, pp. 422–9, 1995.

[123] J. P. Lerch *et al.*, "Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls," *Neurobiology of Aging*, vol. 29, no. 1, pp. 23–30, 2008.

[124] G. Chetelat and J.-C. Baron, "Early diagnosis of alzheimer's disease: contribution of structural neuroimaging," *NeuroImage*, vol. 18, no. 2, pp. 525–541, 2003.

[125] S. Klöppel *et al.*, "Accuracy of dementia diagnosis - A direct comparison between radiologists and a computerized method," *Brain*, vol. 131, no. 11, pp. 2969–2974, 2008.

[126] Department of Health and Social Care, "NHS reference costs 2015 to 2016," Tech. Rep., 2016. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment%7B%5C_%7Ddata/file/577083/Reference%7B%5C_%7DCosts%7B%5C_%7D2015-16.pdf.

[127] M. Boustani, L. Watson, B. Fultz, A. J. Perkins, and R. Druckenbrod, "Acceptance of dementia screening in continuous care retirement communities: A mailed survey," *International Journal of Geriatric Psychiatry*, vol. 18, no. 9, pp. 780–786, 2003.

[128] C. Jagger *et al.*, "Prognosis with dementia in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group," *Neurology*, vol. 54, pp. 16–20, 2000.

[129] P. W. Overstall, "Epidemiology and pathophysiology of falls," in *Fits, Faints and Falls in Old age*, Springer, 1985, pp. 15–26.

[130] C. Van Doorn *et al.*, "Dementia as a Risk Factor for Falls and Fall Injuries Among Nursing Home Residents," *Journal of the American Geriatrics Society*, vol. 51, no. 9, pp. 1213–1218, 2003.

[131] E. Ford *et al.*, "Predicting dementia from primary care records: A systematic review and meta-analysis," *PLOS ONE*, vol. 13, no. 3, G. Forloni, Ed., 2018.

[132] A. Spooner *et al.*, "A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction," *Scientific Reports*, vol. 10, no. 1, p. 20 410, 2020.

[133] P. Battista, C. Salvatore, and I. Castiglioni, "Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: A machine learning study," *Behavioural Neurology*, vol. 1, 2017.

[134] J. Williams and A. Weakley, "Machine Learning Techniques for Diagnostic Differentiation of Mild Cognitive Impairment and Dementia," *AAAI Workshop - Technical Report*, pp. 71–76, 2013.

[135] A. Weakley, J. A. Williams, M. Schmitter-Edgecombe, and D. J. Cook, "Neuropsychological test selection for cognitive impairment classification: A machine learning approach," *Journal of Clinical and Experimental Neuropsychology*, vol. 37, no. 9, pp. 899–916, 2015.

[136] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana, and A. De Mendonça, "Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests," *BMC Research Notes*, vol. 4, no. 1, p. 299, 2011.

[137] P. Vemuri *et al.*, "Alzheimer's disease diagnosis in individual subjects using structural MR images: Validation studies," *NeuroImage*, vol. 39, no. 3, pp. 1186–1197, 2008.

[138] C. Davatzikos, Y. Fan, X. Wu, D. Shen, and S. M. Resnick, "Detection of prodromal Alzheimer's disease via pattern classification of MRI," *Neurobiology of Aging*, vol. 29, no. 4, pp. 514–523, 2009.

[139] Y. Fan, N. Batmanghelich, C. M. Clark, and C. Davatzikos, "Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline," *NeuroImage*, vol. 39, no. 4, pp. 1731–1743, 2008.

[140] S. Duchesne, A. Caroli, C. Geroldi, C. Barillot, G. B. Frisoni, and D. L. Collins, "MRI-based automated computer classification of probable AD versus normal controls," *IEEE Transactions on Medical Imaging*, vol. 27, no. 4, pp. 509–520, 2008.

[141] Z. Lao, D. Shen, Z. Xue, B. Karacali, S. M. Resnick, and C. Davatzikos, "Morphological classification of brains via high-dimensional shape transformations and machine learning methods," *NeuroImage*, vol. 21, no. 1, pp. 46–57, 2004.

[142] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, and J. Tohka, "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects," *NeuroImage*, vol. 104, pp. 398–412, 2015.

[143] A. Abdulkadir, B. Mortamet, P. Vemuri, C. R. Jack, G. Krueger, and S. Klöppel, "Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier," *NeuroImage*, vol. 58, no. 3, pp. 785–792, 2011.

[144] S. G. Costafreda *et al.*, "Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment," *NeuroImage*, vol. 56, no. 1, pp. 212–219, 2011.

[145] C. Davatzikos, S. Resnick, X. Wu, P. Parmpi, and C. Clark, "Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI," *NeuroImage*, vol. 41, no. 4, pp. 1220–1227, 2008.

[146] L. Ferrarini *et al.*, "Ventricular shape biomarkers for Alzheimer's disease in clinical MR images," *Magnetic Resonance in Medicine*, vol. 59, no. 2, pp. 260–267, 2008.

[147] S. Klöppel *et al.*, "Automatic classification of MR scans in Alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, 2008.

[148] S. J. Teipel *et al.*, "Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment," *NeuroImage*, vol. 38, no. 1, pp. 13–24, 2007.

[149] H. Wolf *et al.*, "Structural correlates of mild cognitive impairment," *Neurobiology of Aging*, vol. 25, no. 7, pp. 913–924, 2004.

[150] A. M. Jauhiainen *et al.*, "Discriminating accuracy of medial temporal lobe volumetry and fMRI in mild cognitive impairment," *Hippocampus*, vol. 19, no. 2, pp. 166–175, 2009.

[151] W. Jarrold *et al.*, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 27–37, 2014.

[152] P. Garrard, V. Rentoumi, B. Gesierich, B. Miller, and M. L. Gorno-Tempini, "Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse," *Cortex*, vol. 55, no. 1, pp. 122–129, 2014.

[153] P. Garrard and R. Forsyth, "Abnormal discourse in semantic dementia: A data-driven approach," *Neurocase*, vol. 16, no. 6, pp. 520–528, 2010.

[154] L. A. McGuinness, C. Warren-Gash, L. R. Moorhouse, and S. L. Thomas, "The validity of dementia diagnoses in routinely collected electronic health records in the United Kingdom: A systematic review," *Pharmacoepidemiology and Drug Safety*, vol. 28, no. 2, pp. 244–255, 2019.

[155] Scottish Health Informatics Programme, "SHIP: A Blueprint for Health Records Research in Scotland," Scottish Health Informatics Programme, Tech. Rep., 2012.

[156] C. D. J. Holman, A. J. Bass, I. L. Rouse, and M. S. T. Hobbs, "Population-based linkage of health records in Western Australia: development of a health services research linked database," *Australian and New Zealand Journal of Public Health*, vol. 23, no. 5, pp. 453–459, 1999.

[157] M. J. Schull *et al.*, "ICES: Data, Discovery, Better Health," *International Journal of Population Data Science*, vol. 4, no. 2, 2020.

[158] L. L. Roos, M. Brownell, L. Lix, N. P. Roos, R. Walld, and L. MacWilliam, "From health research to social research: privacy, methods, approaches," *Social Science & Medicine*, vol. 66, no. 1, pp. 117–129, 2008.

[159] S. Greenland, M. A. Mansournia, and D. G. Altman, "Sparse data bias: a problem hiding in plain sight," *British Medical Journal*, vol. 353, p. i1981, 2016.

[160] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[161] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[162] A. J. Mitchell, N. Meader, and M. Pentzek, "Clinical recognition of dementia and cognitive impairment in primary care: a meta-analysis of physician accuracy," *Acta Psychiatrica Scandinavica*, vol. 124, no. 3, pp. 165–183, 2011.

[163] D. Aschwanden *et al.*, "Predicting Cognitive Impairment and Dementia: A Machine Learning Approach," *Journal of Alzheimer's Disease*, vol. 75, no. 3, K. Deckers, Ed., pp. 717–728, 2020.

[164] A. Kluger, S. H. Ferris, J. Golomb, M. S. Mittelman, and B. Reisberg, "Neuropsychological prediction of decline to dementia in nondemented elderly," *Journal of Geriatric Psychiatry and Neurology*, vol. 12, no. 4, pp. 168–179, 1999.

[165] E. A. Phelan, S. Borson, L. Grothaus, S. Balch, and E. B. Larson, "Association of Incident Dementia With Hospitalizations," *Journal of the American Medical Association*, vol. 307, no. 2, p. 165, 2012.

[166] S. Toot, M. Devine, A. Akporobaro, and M. Orrell, "Causes of Hospital Admission for People With Dementia: A Systematic Review and Meta-Analysis," *Journal of the American Medical Directors Association*, vol. 14, no. 7, pp. 463–470, 2013.

[167] M. E. Soto, S. Andrieu, S. Gillette-Guyonnet, C. Cantet, F. Nourhashemi, and B. Vellas, "Risk factors for functional decline and institutionalisation among community-dwelling older adults with mild to severe Alzheimer's disease: one year of follow-up," *Age and Ageing*, vol. 35, no. 3, pp. 308–310, 2006.

[168] A. Padkin, C. Goldfrad, A. R. Brady, D. Young, N. Black, and K. Rowan, "Epidemiology of severe sepsis occurring in the first 24 hrs in intensive care units in England, Wales, and Northern Ireland," *Critical Care Medicine*, vol. 31, no. 9, 2003.

[169] C. Fleischmann *et al.*, "Assessment of Global Incidence and Mortality of Hospital-treated Sepsis. Current Estimates and Limitations," *American Journal of Respiratory and Critical Care Medicine*, vol. 193, no. 3, pp. 259–272, 2016.

[170] M. Singer *et al.*, "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *Journal of the American Medical Association*, vol. 315, no. 8, p. 801, 2016.

[171] E. C. Bishop, *Early Identification and Treatment of Sepsis: Clinical Guideline*, 2017. [Online]. Available: `https://www.meht.nhs.uk/EasysiteWeb/getresource.axd?AssetID=18015%7B%5C&%7Dtype=full%7B%5C&%7Dservicetype=Attachment`.

[172] R. Daniels, "Surviving the first hours in sepsis: getting the basics right (an intensivist's perspective)," *Journal of Antimicrobial Chemotherapy*, vol. 66, no. 2, pp. 11–23, 2011.

[173] R. A. Balk, "Systemic inflammatory response syndrome (SIRS): Where did it come from and is it still relevant today?" *Virulence*, vol. 5, no. 1, pp. 20–26, 2014.

[174] R. C. Bone *et al.*, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis," *Chest*, vol. 101, no. 6, pp. 1644–1655, 1992.

[175] J. Hajj, N. Blaine, J. Salavaci, and D. Jacoby, "The "Centrality of Sepsis": A Review on Incidence, Mortality, and Cost of Care," *Healthcare*, vol. 6, no. 3, p. 90, 2018.

[176] C. W. Seymour *et al.*, "Time to treatment and mortality during mandated emergency care for sepsis," *New England Journal of Medicine*, vol. 376, no. 23, pp. 2235–2244, 2017.

[177] R. Serafim, J. A. Gomes, J. Salluh, and P. Póvoa, "A Comparison of the Quick-SOFA and Systemic Inflammatory Response Syndrome Criteria for the Diagnosis of Sepsis and Prediction of Mortality: A Systematic Review and Meta-Analysis.," *Chest*, vol. 153, no. 3, pp. 646–655, 2018.

[178] L. M. Fleuren *et al.*, "Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy," *Intensive Care Medicine*, vol. 46, no. 3, pp. 383–400, 2020.

[179] T. Desautels *et al.*, "Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach," *JMIR Medical Informatics*, vol. 4, no. 3, p. 28, 2016.

[180] A. J. Masino *et al.*, "Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data," *PLOS ONE*, vol. 14, no. 2, J. M. Juarez, Ed., 2019.

[181] S. Le *et al.*, "Pediatric Severe Sepsis Prediction Using Machine Learning," *Frontiers in Pediatrics*, vol. 7, no. 10, pp. 1–8, 2019.

[182] D. W. Shimabukuro, C. W. Barton, M. D. Feldman, S. J. Mataraso, and R. Das, "Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial," *BMJ Open Respiratory Research*, vol. 4, no. 1, 2017.

[183] J. S. Calvert *et al.*, "A computational approach to early sepsis detection," *Computers in Biology and Medicine*, vol. 74, pp. 69–73, 2016.

[184] H. J. Kam and H. Y. Kim, "Learning representations for the early detection of sepsis with deep neural networks," *Computers in Biology and Medicine*, vol. 89, no. April, pp. 248–255, 2017.

[185] M. Faisal *et al.*, "Development and External Validation of an Automated Computer-Aided Risk Score for Predicting Sepsis in Emergency Medical Admissions Using the Patient's First Electronically Recorded Vital Signs and Blood Test Results*," *Critical Care Medicine*, vol. 46, no. 4, pp. 612–618, 2018.

[186] S. Horng, D. A. Sontag, Y. Halpern, Y. Jernite, N. I. Shapiro, and L. A. Nathanson, "Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning," *PLOS ONE*, vol. 12, no. 4, T. Groza, Ed., e0174708, 2017.

[187] F. van Wyk, A. Khojandi, A. Mohammed, E. Begoli, R. L. Davis, and R. Kamaleswaran, "A minimal set of physiomarkers in continuous high frequency data streams predict adult sepsis onset earlier," *International Journal of Medical Informatics*, vol. 122, no. 1, pp. 55–62, 2019.

[188] L. M. Kalisch Ellett, N. L. Pratt, E. N. Ramsay, J. D. Barratt, and E. E. Roughead, "Multiple anticholinergic medication use and risk of hospital admission for confusion or dementia," *Journal of the American Geriatrics Society*, vol. 62, no. 10, pp. 1916–1922, 2014.

[189] M. Chan, F. Nicklason, and J. H. Vial, "Adverse drug events as a cause of hospital admission in the elderly," *Internal Medicine Journal*, vol. 31, no. 4, pp. 199–205, 2001.

[190] A. Natalwala, R. Potluri, H. Uppal, and R. Heun, "Reasons for Hospital Admissions in Dementia Patients in Birmingham, UK, during 2002–2007," *Dementia and Geriatric Cognitive Disorders*, vol. 26, no. 6, pp. 499–505, 2008.

[191] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot Ensembles: Train 1, get M for free," pp. 1–14, 2017.

[192] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," *International Conference on Learning Representations*, pp. 1–16, 2016.

[193] D. V. Ford *et al.*, "The SAIL Databank: building a national architecture for e-health research and evaluation.," *BMC health services research*, vol. 9, no. 1, p. 157, 2009.

[194] M. O'neil, C. Payne, and J. Read, "Read Codes Version 3: a user led terminology," *Methods of Information in Medicine*, vol. 34, pp. 187–192, 1995.

[195] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[196] D. D. Tresch, M. F. Folstein, P. V. Rabins, and W. R. Hazzard, "Prevalence and Significance of Cardiovascular Disease and Hypertension in Elderly Patients With Dementia and Depression," *Journal of the American Geriatrics Society*, vol. 33, no. 8, pp. 530–537, 1985.

[197] M. Sanderson, J. Wang, D. R. Davis, M. J. Lane, C. B. Cornman, and M. K. Fadden, "Co-morbidity associated with dementia," *American Journal of Alzheimer's Disease & Other Dementiasr*, vol. 17, no. 2, pp. 73–78, 2002.

[198] H. K. Kamel, M. S. Hussain, S. Tariq, H. M. Perry, and J. E. Morley, "Failure to diagnose and treat osteoporosis in elderly patients hospitalized with hip fracture," *The American Journal of Medicine*, vol. 109, no. 4, pp. 326–328, 2000.

[199] C. Feart *et al.*, "Associations of lower vitamin D concentrations with cognitive decline and long-term risk of dementia and Alzheimer's disease in older adults," *Alzheimer's & Dementia*, vol. 13, no. 11, pp. 1207–1216, 2017.

[200] Y. Sato, J. Iwamoto, T. Kanoko, and K. Satoh, "Amelioration of Osteoporosis and Hypovitaminosis D by Sunlight Exposure in Hospitalized, Elderly Women With Alzheimer's Disease: A Randomized Controlled Trial," *Journal of Bone and Mineral Research*, vol. 20, no. 8, pp. 1327–1333, 2005.

[201] P. L. Gozalo, A. Pop-Vicas, Z. Feng, S. Gravenstein, and V. Mor, "Effect of Influenza on Functional Decline," *Journal of the American Geriatrics Society*, vol. 60, no. 7, pp. 1260–1267, 2012.

[202] Royal Pharmaceutical Society of Great Britain, *British National Formulary 58*. Royal Pharmaceutical Society, 2009.

[203] R. Bitton, "The economic burden of osteoarthritis.," *The American journal of managed care*, vol. 15, no. 8, pp. 230–235, 2009.

[204] I. E. van de Vorst, H. L. Koek, C. E. Stein, M. L. Bots, and I. Vaartjes, "Socioeconomic Disparities and Mortality After a Diagnosis of Dementia: Results From a Nationwide Registry Linkage Study," *American Journal of Epidemiology*, vol. 184, no. 3, pp. 219–226, 2016.

[205] P. Voyer, S. Richard, L. Doucet, C. Danjou, and P. H. Carmichael, "Detection of delirium by nurses among long-term care residents with dementia," *BMC Nursing*, vol. 7, pp. 1–14, 2008.

[206] M. A. Reyna *et al.*, "Early Prediction of Sepsis From Clinical Data," *Critical Care Medicine*, vol. 48, no. 2, pp. 210–217, 2020.

[207] C. Brun-Buisson, F. Roudot-Thoraval, E. Girou, C. Grenier-Sennelier, and I. Durand-Zaleski, "The costs of septic syndromes in the intensive care unit and influence of hospital-acquired sepsis," *Intensive Care Medicine*, vol. 29, no. 9, pp. 1464–1471, 2003.

[208] D. B. Page, J. P. Donnelly, and H. E. Wang, "Community-, Healthcare-, and Hospital-Acquired Severe Sepsis Hospitalizations in the University HealthSystem Consortium," *Critical Care Medicine*, vol. 43, no. 9, pp. 1945–1951, 2015.

[209] F. O. Odetola, A. Gebremariam, and G. L. Freed, "Patient and hospital correlates of clinical outcomes and resource utilization in severe pediatric sepsis," *Pediatrics*, vol. 119, no. 3, pp. 487–494, 2007.

[210] H. E. Wang, N. I. Shapiro, D. C. Angus, and D. M. Yealy, "National estimates of severe sepsis in United States emergency departments," *Critical Care Medicine*, vol. 35, no. 8, pp. 1928–1936, 2007.

[211] C. J. Paoli, M. A. Reynolds, M. Sinha, M. Gitlin, and E. Crouser, "Epidemiology and Costs of Sepsis in the United States—An Analysis Based on Timing of Diagnosis and Severity Level," *Critical Care Medicine*, vol. 46, no. 12, pp. 1889–1897, 2018.

[212] R. Barandela, R. M. Valdovinos, J. S. Sánchez, and F. J. Ferri, "The Imbalanced Training Sample Problem: Under or over Sampling?" In *Lecture Notes in Computer Science*, 2004, pp. 806–814.

[213] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman, "An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU," *Critical Care Medicine*, vol. 46, no. 4, pp. 547–553, 2018.

[214] G. C. Siontis, I. Tzoulaki, P. J. Castaldi, and J. P. Ioannidis, "External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination," *Journal of Clinical Epidemiology*, vol. 68, no. 1, pp. 25–34, 2015.

[215] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations*, pp. 1–15, 2014.

[216] L. Keeley, "Reducing the risk of ventilator-acquired pneumonia through head of bed elevation," *Nursing in Critical Care*, vol. 12, no. 6, pp. 287–294, 2007.

[217] N. A. Metheny and R. A. Frantz, "Head-of-Bed Elevation in Critically Ill Patients: A Review," *Critical Care Nurse*, vol. 33, no. 3, pp. 53–67, 2013.

[218] E. D. P. Afonso and S. Blot, "Effect of gestational age on the epidemiology of late-onset sepsis in neonatal intensive care units-a review," *Expert Review of Anti-Infective Therapy*, vol. 15, no. 10, pp. 917–924, 2017.

[219] A. Belachew and T. Tewabe, "Neonatal sepsis and its association with birth weight and gestational age among admitted neonates in Ethiopia: systematic review and meta-analysis," *BMC Pediatrics*, vol. 20, no. 1, p. 55, 2020.

[220] Y. Futagi, Y. Toribe, and Y. Suzuki, "The Grasp Reflex and Moro Reflex in Infants: Hierarchy of Primitive Reflex Responses," *International Journal of Pediatrics*, vol. 2012, pp. 1–10, 2012.

[221] K. D. Fairchild *et al.*, "Vital signs and their cross-correlation in sepsis and NEC: a study of 1,065 very-low-birth-weight infants in two NICUs," *Pediatric Research*, vol. 81, no. 2, pp. 315–321, 2017.

[222] B. J. Stoll *et al.*, "Very low birth weight preterm infants with early onset neonatal sepsis: the predominance of gram-negative infections continues in the National Institute of Child Health and Human Development Neonatal Research Network, 2002–2003," *The Pediatric Infectious Disease Journal*, vol. 24, no. 7, pp. 635–639, 2005.

[223] G. M. Savva, S. B. Wharton, P. G. Ince, G. Forster, F. E. Matthews, and C. Brayne, "Age, Neuropathology, and Dementia," *New England Journal of Medicine*, vol. 360, no. 22, pp. 2302–2309, 2009.

[224] Joint Formulary Committee, Royal Pharmaceutical Society of Great Britain, *British national formulary*. Pharmaceutical Press, 2020, vol. 80.

[225] A. Hofman *et al.*, "Atherosclerosis, apolipoprotein E, and prevalence of dementia and Alzheimer's disease in the Rotterdam Study," *The Lancet*, vol. 349, no. 9046, pp. 151–154, 1997.

[226] M. Nagai, S. Hoshide, and K. Kario, "Hypertension and Dementia," *American Journal of Hypertension*, vol. 23, no. 2, pp. 116–124, 2010.

[227] K. M. de Almondes, M. V. Costa, L. F. Malloy-Diniz, and B. S. Diniz, "Insomnia and risk of dementia in older adults: Systematic review and meta-analysis," *Journal of Psychiatric Research*, vol. 77, pp. 109–115, 2016.

[228] P. Lass, J. R. Buscombe, M. Harber, A. Davenport, and A. J. W. Hilson, "Cognitive impairment in patients with renal failure is associated with multiple-infarct dementia," *Clinical Nuclear Medicine*, vol. 24, no. 8, pp. 561–565, 1999.

[229] A. L. Byers and K. Yaffe, "Depression and risk of developing dementia," *Nature Reviews Neurology*, vol. 7, no. 6, pp. 323–331, 2011.

[230] J. Noebels, "A perfect storm: Converging paths of epilepsy and Alzheimer's dementia intersect in the hippocampal formation," *Epilepsia*, vol. 52, no. 1, pp. 39–46, 2011.

[231] I. M. Rosa, A. G. Henriques, L. Carvalho, J. Oliveira, and O. A. B. d. C. e Silva, "Screening younger individuals in a primary care setting flags putative dementia cases and correlates gastrointestinal diseases with poor cognitive performance," *Dementia and Geriatric Cognitive Disorders*, vol. 43, no. 1-2, pp. 15–28, 2017.

[232] X. Hu, Z. Lu, Z. Mao, and J. Yin, "Association between myasthenia gravis and cognitive function: A systematic review and meta-analysis," *Annals of Indian Academy of Neurology*, vol. 18, no. 2, p. 131, 2015.

[233] A. J. Mitchell, "A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment," *Journal of Psychiatric Research*, vol. 43, no. 4, pp. 411–431, 2009.

[234] C. A. de Jager, A.-C. M. C. Schrijnemaekers, T. E. M. Honey, and M. M. Budge, "Detection of MCI in the clinic: evaluation of the sensitivity and specificity of a computerised test battery, the Hopkins Verbal Learning Test and the MMSE," *Age and Ageing*, vol. 38, no. 4, pp. 455–460, 2009.

[235] K. Schultz-Larsen, R. K. Lomholt, and S. Kreiner, "Mini-Mental Status Examination: A short form of MMSE was as accurate as the original MMSE in predicting dementia," *Journal of Clinical Epidemiology*, vol. 60, no. 3, pp. 260–267, 2007.

[236] A. Rusanov, N. G. Weiskopf, S. Wang, and C. Weng, "Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research," *BMC Medical Informatics and Decision Making*, vol. 14, no. 1, p. 51, 2014.

[237] J. B. Lee, R. A. Rossi, S. Kim, N. K. Ahmed, and E. Koh, "Attention Models in Graphs," *ACM Transactions on Knowledge Discovery from Data*, vol. 13, no. 6, pp. 1–25, 2019.

[238] D. Hu, "An introductory survey on attention mechanisms in NLP problems," *Advances in Intelligent Systems and Computing*, vol. 1038, pp. 432–448, 2020.

[239] M.-H. Guo *et al.*, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, 2022.

[240] J. Yoon, J. Jordon, and M. Van Der Schaar, "Invase: Instance-wise variable selection using neural networks," *International Conference on Learning Representations*, pp. 1–24, 2019.