Can Machine Learning be used as an advanced diagnostic tool in Diabetes Mellitus?

MSc Stratified Medicine and Pharmacological Innovation

Student Number: 2023061
Student Name: Ruth Muir

Project Supervisor: Dr Christopher Sainsbury

Organisation: NHS Greater Glasgow and Clyde - Diabetology

ABSTRACT:

Introduction: The science of Artificial Intelligence is rapidly expanding into all fields of medicinal research. By utilising supervised and unsupervised models of Machine Learning, a more advanced diagnostic model could be developed to aid physicians in diabetes diagnosis through the development of Artificial Neural Networks(ANN).
Methods: NHS Greater Glasgow and Clyde SCI-Diabetes database was utilised as the patient database. 6 consultant diabetologists (physicians) were given datasets of 100 patients and asked to diagnosis Type 1 or Type 2 diabetes mellitus from the provided parameters. A logistic regression and ANN were also undertaken on each physician dataset, and on the NHSGGC SCI-Diabetes dataset as a whole. Diagnostic accuracy of each model was then calculated from current patient diagnosis as recorded on the SCI diabetes system.
Results: The ANN model was the most accurate approach to diabetes diagnosis, whilst the physician opinion was concluded the least accurate. Both the ANN and logistic regression were shown to be excellent models at positively predicting correct diagnosis with AUC values of 0.95.
Discussion: Due to the lack of consistency between the ANN, logistic regression, physician diagnosis, and the current diagnosis of a number of patients, there may be an issue of data quality. This issue of misdiagnosis could be solved by combining the supervised ANN model with an unsupervised SOM approach to highlight patients with the greatest probability of misdiagnosis.
Conclusion: The utilisation of machine learning in diabetes diagnosis to aid in physician prediction could result in the implementation of a highly accurate diagnostic system and assist in highlighting individuals that required diagnosis re-assessment.

INTRODUCTION

Artificial Intelligence

Artificial Intelligence(AI) is the science of utilising computers to understand human intelligence and to adapt a computational system to achieve intellectual properties, based on studying of the conventional neurological systems. The aim of AI is to develop computational programmes that can learn, think and behave like a human to achieve specified goals, or make predictions from complex data that would be impossible for a human being to conclude (McCarthy, 2001).

The field of AI is rapidly expanding, and is a key component of all large social network, commercial and data aggregation platforms such as Amazon, Google and Microsoft. Due to the increasing amount of digital data that is collected on a daily basis, there is a major requirement for the development of technology to efficiently utilise this data, and process it in an efficient manner. Recent achievements in the field of AI include the development of text analytics to combat the issue of large data accumulation and to allow companies to interpret data in a way that would be physically impossible without the use of computers. An example of this can be found in the field of hospitality, where companies are now utilising text analytic programmes to

determine overall customer satisfaction and highlighting key areas of concern (Xiang et al, 2015).

Machine Learning

Machine Learning is a type of AI that utilises the machine to independently complete tasks on new, and complex datasets. Machine learning can be utilised as a data mining tool to extract useful information from large volumes of data, or it can be used to develop a probabilistic predictive power when presented with new datasets (Maglogiannis, 2007). Machine learning can be further categorised, with supervised and unsupervised models two of the main subcategories. Supervised machine learning models generalise outcomes once exposed to an initial training set of data. The computer can adapt the algorithm developed from the training set to learn from the data for future decision making. Artificial Neural Networks (ANN) are a type of supervised deep learning, which depend on the architecture of the network developed and are built by a series of input and activation units, outputs, cycle repetition and factor weighting. ANNs utilise the training set to enhance the predictive power of outcome, which can then be used on original data to accurately predict a diagnosis given defined parameters. The parameters selected should allow the development of a flexible model, to allow self-regulation of the predictive model to develop a more accurate predictive power than in the presence of limited parameters or parameters unrelated to the diagnostic problem (Ghahramani, 2015).

Unsupervised machine learning allows the network to draws its own conclusions from data, without an initial training set to input classifications. Unsupervised methods are most often utilised to draw unidentified associations within the data, and to identify clustering (Hodeghatta and Nayak, 2017). Self-Organising Maps (SOM) are one of the most common types of ANN developed by an unsupervised approach. SOMs are a two-dimensional topological representation of input data, developed using competitive learning (Kohonen,1982). By utilising a combination of machine learning methods, it is hoped that an efficient classification model could be developed to increase diagnostic accuracy over current healthcare systems.

AI and Medicine

Although a leading contributor in the advancement of a number of fields, AI is only beginning to be effectively utilised in medicine and the field of healthcare. Recently, researchers have shown than everyday CT images can be utilised to predict patient longevity using machine learning neural network computer analysis of radiomic biomarkers. ANN analysis of patients CT scans have shown to be as effective as many more complex and expensive mortality analysis methods, and so is praised as a cost-efficient method of precision medicine, as treatment can then be offered further on a patient by patient basis (Oakden-Rayner et al., 2017). Although advancements are now beginning to occur, there is still a long way to go for AI to reach full potential in the healthcare setting. If integrated efficiently, AI would lead to an expansion of life-saving research opportunities and expand diagnostic possibilities.

Diabetes Mellitus

Increased diagnostic accuracy could be important in the development of diabetic healthcare, ensuring patients are diagnosed correctly to increase treatment efficiency and decrease cost of incorrect treatment methods. In Scotland, Diabetes Mellitus would be a potential benefactor from machine learning development due to the vast quantities of data held by each diabetic patient by SCI- Diabetes. SCI-Diabetes is a national dataset of all patients with a primary care code for diabetes, and is one of the most complete national diabetes dataset globally, using Scotland as a test model for the potential use of such an efficient dataset collection. There are a variety of national diabetes datasets developed by other countries worldwide, and approaches developed using SCI-Diabetes have the potential to be more widely adopted.SCI-Diabetes links data from other national repositories such as mortality, biochemical and general health results, and has accumulated up to 25 years of patient data from some individuals. The database is currently used for clinical practice, as a method of record keeping and to assist clinical risk stratification. By utilising the data available from SCI-Diabetes, machine learning could be used as an efficient tool to aid in diagnostic accuracy.

To test the initial effectiveness of utilising the SCI-diabetes data for diagnostic purposes, Type 1 and Type 2 diabetes patients were selected. By training a model to effectively distinguish between the two main diabetes types, it would then be possible to expand the model scope, and include all diabetic patient types to increase accuracy of diagnosis over current methods.

The aim of the project was to develop an Artificial Neural Network(ANN) with a diabetic diagnostic accuracy approaching current classical computerised approaches, and to test this approach against clinical judgement - ie the diagnostic classification chosen by physicians specialising in diabetes when presented with the same initial data. A secondary aim was to develop a method of data comparison that would allow the identification of patients with a high probability of misdiagnosis.

METHODS

Dataset Refinement

The NHS Greater Glasgow and Clyde SCI-Diabetes database was utilised to obtain an initial basis for the dataset which would be further tested upon. The database included every patient within the Greater Glasgow and Clyde region to have been diagnosed with Diabetes, and all of their medical information that has been linked to the database. The broad dataset was refined and cleaned to allow patients with the required parameters to be included. Statistical computing software packages R (R core team, 2017) and Python Software Foundation (Python Language Reference, version 3.4) were used to develop the ANN and logistic regression, as well as results collected from each model. Patients with missing data were removed as project criteria inhibited imputation methods for data entry to be utilised. Furthermore, patients who did not have a Type 1 or Type 2 diabetes diagnosis were removed from the dataset as the primary

focus of the project required binary classification. Finally, date of diagnosis was analysed, and patients with a diagnosis interval of less than 1 year were removed from the dataset to increase confidence in accuracy of recorded diagnosis.

Data from the remaining patients was merged with ID files to include all of the required parameters; BMI, systolic BP, diastolic BP, HbA1c levels, Date of Birth and gender. As each individual had multiple values for a number of examinable fields, a method of value selection was developed. To obtain the most accurate systolic BP value for each individual, a 3 month window prior and following date of diagnosis was set. Systolic BP values within the time frame were then plotted in relation to time and squared/ square rooted to remove negative values. The BP value with the lowest time association was then selected for each individual in order to accurately select the value with the highest association with date of diagnosis. This method of value selection was then repeated for diastolic BP, BMI and Hba1c levels. Ethnicity and gender were part of the original dataset, and age (at the time of diagnosis) was calculated from patient Date of Birth to give all required analysable data. This method of data refinement and cleaning determined the cohort of individuals utilised.

Physician Opinion

3 Physicians (diabetologists) were allocated forms which contained the 6 parameters for 100 diabetic patients, and were asked to determine the patient diagnosis given the information provided. Physicians were advised that diagnosis could only be Type 1 or Type 2, and diagnosis ratio for Type 1 to Type 2 differed from real world cases. Physician diagnosis was then directly compared to current patient diagnosis and accuracy assessed.

Logistic Regression and Supervised Machine Learning

A logistic regression was carried out on the entire dataset, to determine accuracy of the classical model compared to current patient diagnosis, and was further carried out on each of the allocated forms given to the physician to determine if the classical approach assessed the same diagnosis as the physician. A supervised machine learning model was then undertaken to assess accuracy of the advanced ANN compared to current patient diagnosis and the logistic regression. As undertaken for the logistic regression, the ANN was carried out on the entire dataset and on each individual allocated physician form.

For both the logistic regression and ANN, the dataset was split into a training and test set in a 80:20 ratio, and both sets were feature scaled to standardise the independent variable values to avoid unnecessary influence of numerically larger variables on diagnosis outcome. A training set was generated to develop an efficient model which was then carried out on the test set to determine diagnosis for patients previously unseen by the training set model. The data allocated to physicians for classification were drawn from the test sets. A prediction threshold value was calculated, and individuals with a prediction value of greater than the threshold were classified as Type 1, with all other patients classified as Type 2.

A confusion matrix was developed with accuracy, sensitivity and specificity values extracted. A ROC curve was plotted and the Area Under the Curve (AUC) was calculated. The equation for sensitivity (true positive rate), specificity (true negative rate) and false positive rate are as follows:

$$\textbf{Sensitivity(True positive rate)} = \frac{\textbf{Correct Type 1 prediction}}{\textbf{Overall no. Type 1 prediction}}$$

$$\textbf{Specificity (True negative rate)} = \frac{\textbf{Correct Type 2 predictions}}{\textbf{Overall no. Type 2 predictions}}$$

$$\textbf{False positive rate} = \frac{\textbf{False Type 1 prediction}}{\textbf{False Type 1 prediction + True negative}}$$

Relative performance of each model and the physician were plotted to visually demonstrate superiority of the differing methods.

ANN model dimensions

The ANN model developed was a 2 layer input comprised of 16 neurons units per layer, and 1 output. The model was set to run 50 epoch cycles. After each cycle the model was optimised by comparing results of the training set to the actual patient diagnosis. This optimisation allowed the refinement of the model relating to the independent variables and their weighting of importance to the overall diagnosis. This process allowed continuous adaption of the model when faced with a wide range of individuals, and increased accuracy when unseen patients were encountered in the test set.

Error Analysis

Results were collected from each model and compared to current patient diagnosis. Error analysis was then undertaken in those cases where model determined diagnosis differed from the current diagnosis within SCI-Diabetes.

RESULTS

Statistical Analysis of Cohort

Prior to ANN and logistic regression data manipulation, a descriptive statistical analysis was carried out on the cohort of patients within the dataset to determine baseline values of each examinable factor.

Table 1: Statistical Summary

Factors	General patient cohort	Type 1 patients	Type 2 patients
Cohort number	49,995	3,222	46,773
Sex	Female: 22,124 Male: 27,871	Female: 1,388 Male: 1,834	Female: 20,736 Male: 26,037
Mean Age (years)	56.7 (48.0, 67.2)	28.5 (15.1, 39.9)	58.6 (49.8, 67.8)
Mean BMI	31.8 (27.2, 35.6)	23.4 (19.5, 26.4)	32.4 (27.8, 35.9)
Mean hba1c	66.0 (49.0, 79.0)	86.1 (63.0, 107.0)	64.9 (49.0, 77.0)
Mean sbp	136.5 (124.0, 147.0)	120.8 (110.0, 130.0)	137.5 (125.0, 148.0)
Mean dbp	80.2 (72.0, 87.0)	72.6 (63.0, 80.0)	80.7 (73.0, 88.0)

Table 1 illustrates the demographics of the input cohorts, comparing the general cohort to both Type 1 and Type 2 patients.

Table 1 shows that 6% of the patients within the dataset were diagnosed with Type 1 diabetes. The gender ratio remained equal across all groups, at around 56% of each cohort male. Mean BMI was 28% lower in Type 1 patients at 23.4, compared to 32.4 of Type 2 patients. Mean hba1c value also fluctuated between diagnosis Type, with Type 2 patients mean hba1c value 25% lower at 64.9 compared to 86.1 of Type 1 individuals. Finally, mean systolic and diastolic BP varied between the groups, with the average BP of Type 1 patients 121/ 73 and Type 2 individuals with an overall higher BP of 138/ 81.

Table 2: Ethnicity

Ethnicity	General patient cohort	Type 1 patients	Type 2 patients	Percentage of Type 1 Patients
African	297	30	267	10.1%
Bangladeshi	353	11	342	3.1%
Chinese	246	11	235	4.5%
Indian	743	21	722	2.8%
White - British	8,281	384	7,897	4.6%
White - Scottish	26,466	1,732	24,734	6.5%

Ethnicity	General patient cohort	Type 1 patients	Type 2 patients	Percentage of Type 1 Patients
Multiple Ethnicity	499	23	476	4.6%
Pakistani	1,187	37	1,150	3.1%
Other - Asian	451	17	434	3.8%
Other - white	1,460	122	1,338	8.3%
Other - other ethnic	10,012	834	9,178	8.3%

Table 2 illustrates the ethnic groups present within the cohort.

Table 2 demonstrates the ethnic groups within the general cohort, and the percentage of each ethnicity currently diagnosed as Type 1 and Type 2. Percentage of Type 1 patients per ethnicity illustrates that African, Other-white and Other - other ethnic are the most likely to be currently diagnosed with Type 1 diabetes, whilst Indian, Pakistani and Bangladeshi are likely to have Type 2 diabetes.

Confusion Matrix

For both the ANN and logistic regression, a prediction threshold value of 0.1 was calculated, and individuals with a prediction value of > 0.1 were classified as Type 1, with all other patients classified as Type 2. A confusion matrix was generated for each dataset by each model, with the main findings summarised in table 3.

Table 3: Confusion Matrix

	Physician	ANN	Logistic Regression
Accuracy (CI 0.95)	0.86 (0.78,0.92)	0.93 (0.85, 0.96)	0.91 (0.83, 0.96)
Specificity	0.77	0.85	0.81
Sensitivity	0.93	0.96	0.95

Table 3 illustrates the median confusion matrix values collected from the 6 physician forms analysed. The table details accuracy(CI 0.95), specificity and sensitivity for each model which was then utilised in the development of the ROC curve analysis.

The confusion matrix results demostrate that both the classical and advanced models are more efficient in accuracy, specificity and sensitivity when compared to the physician's diagnosis.

When comparing the logistic regression and ANN, the advanced machine learning model illustrates a minor improvement in all fields, however results cannot be considered significantly different.

ROC Curve Analysis

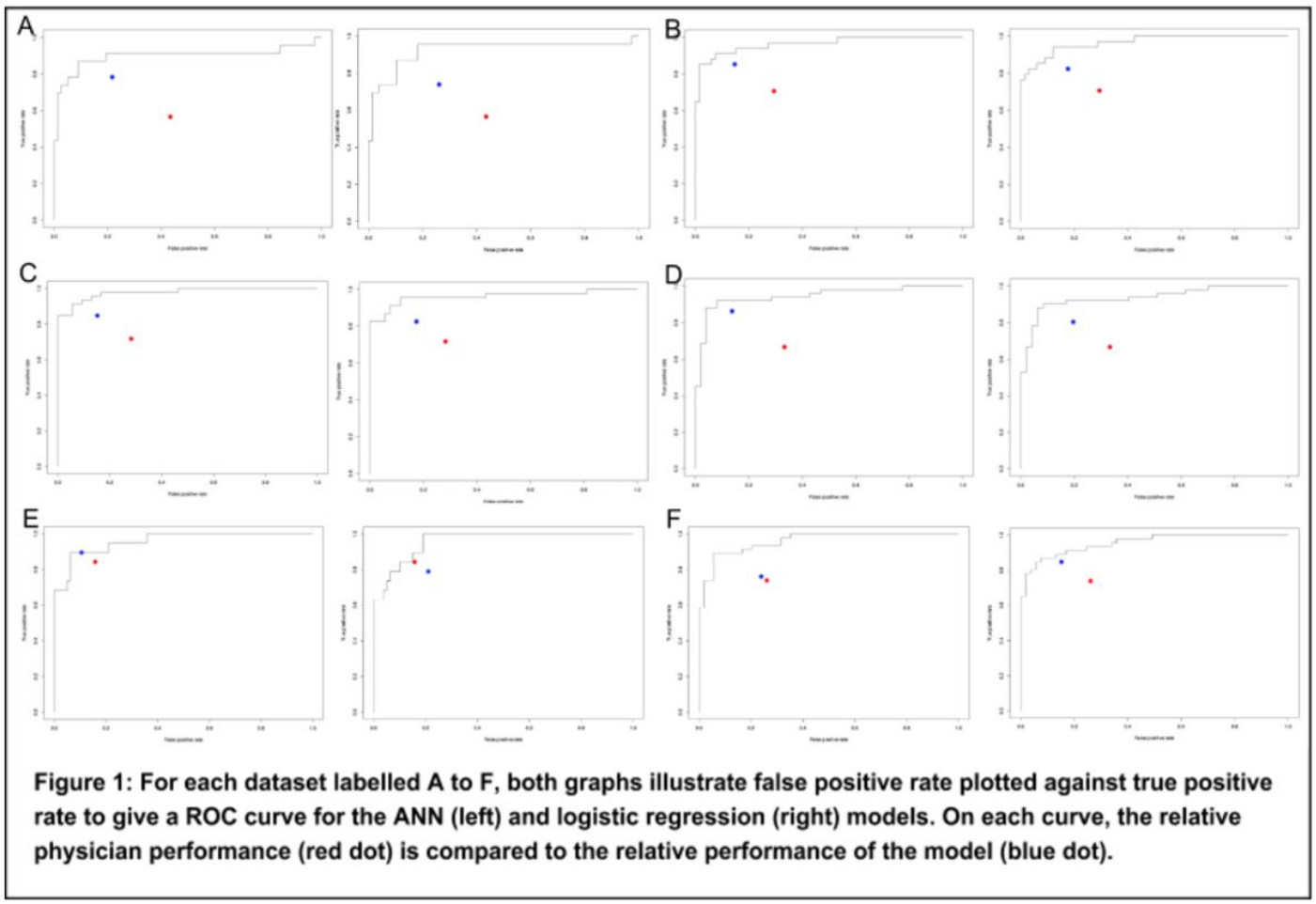

Figure 1: For each dataset labelled A to F, both graphs illustrate false positive rate plotted against true positive rate to give a ROC curve for the ANN (left) and logistic regression (right) models. On each curve, the relative physician performance (red dot) is compared to the relative performance of the model (blue dot).

A ROC curve analysis of each physician form dataset was analysed, comparing false positive and true positive rates, as illustrated in Figure 1. From the ROC curves, Area Under the ROC Curve (AUC) values were calculated for each model. AUC values for ANN ranged from 0.90 to 0.99, with logistic regression values ranging from 0.92 to 0.99. This illustrates that the AUC values calculated remained similar between models throughout the datasets, and that both model AUC values were consistently considered excellent diagnosis predictors (AUC > 0.9).

Also plotted on Figure 1 are the relative performances of the physician and the model per dataset. From dataset A to F, the ANN model is shown on the left with the logistic regression comparison on the right. All ANN graphs illustrate a more efficient relative performance in comparison to the physician, with the model marker to the left and higher than the physician marker. This illustrates that the logistic regression and ANN are illustrating greater true positive,

and lower false positive rates than that of the physician. Datasets A and E illustrate the largest improvement of the model compared to the physician, with the largest marker gap, whereas dataset F shows a minor margin between ANN and physician illustrating the least improvement from the physician by the ANN model. The logistic regression follows the same pattern as the ANN, with 5 of the 6 graphs illustrate a more efficient relative performance of the logistic regression marker compared to that of the physician, however dataset E shows that the physician has outperformed the model. Nevertheless, the overall average performance of the logistic regression model is clearly illustrated as more efficient that the physician alone.

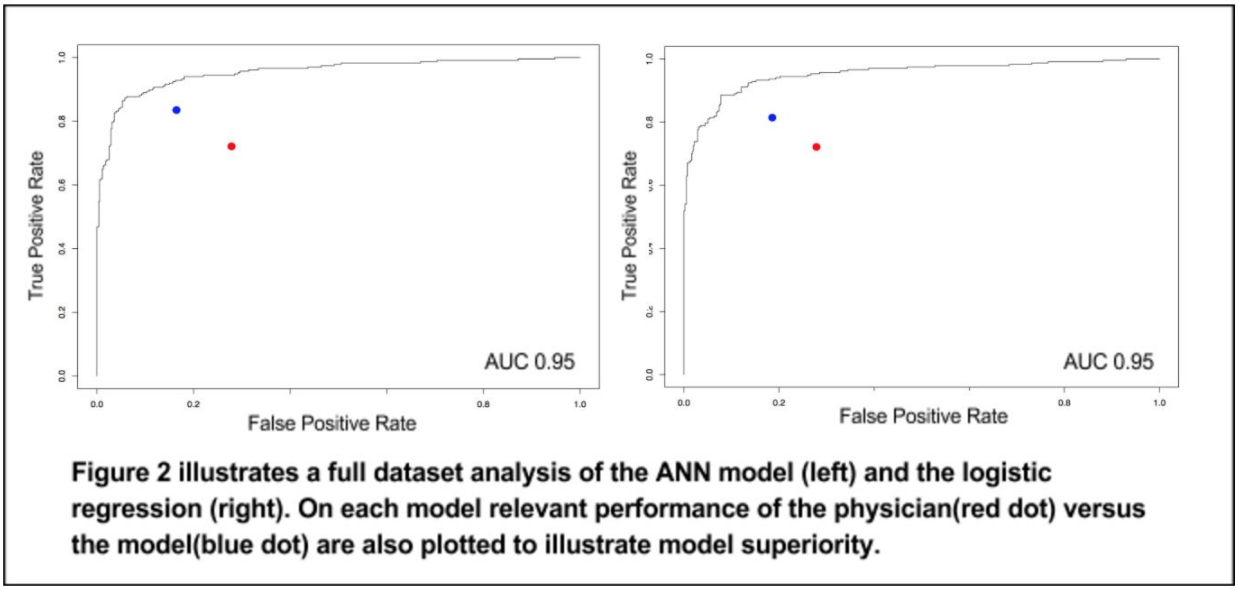

Figure 2 illustrates a full dataset analysis of the ANN model (left) and the logistic regression (right). On each model relevant performance of the physician(red dot) versus the model(blue dot) are also plotted to illustrate model superiority.

Figure 2 illustrates the ROC curve analysis for both the ANN and logistic regression on the full SCI-Diabetes refined cohort. As illustrated, the physician has shown to less efficient at diagnosing Type 1 and Type 2 to that of the current clinical diagnosis, compared to both models. The average AUC for both the ANN and logistic regression was 0.95 demonstrating the similarity between model accuracy. With an AUC of 0.95, both models can be considered excellent at positively predicting the correct diagnosis.

Error Analysis and Model Concordance

Model concordance illustrates that the ANN and logistic regression models were similarly structured with a diagnosis rate of 97% in agreement. This decreased when each model was compared to physician diagnosis, with physician and ANN similarity of 90% and physician and logistic regression at 88% diagnosis similarity. When ANN and physician diagnosis were further compared, the ANN was continuously noted to over-diagnosis Type 1, whereas physicians over-diagnosed Type 2. Type 1 diagnosis in Type 2 patients is less clinically important than Type 1 patients diagnosed as Type 2 therefore ANN diagnosis structure can be considered clinically superior. ANN and current clinical diagnosis were 93% in concordance. This data could

be utilised further clinically if logistic regression and physician diagnosis were also examined. A proportion of patients could then be highlighted that could benefit from diagnosis re-assessment if ANN, logistic regression and physician all disagree with current clinical diagnosis.

ANN model prediction

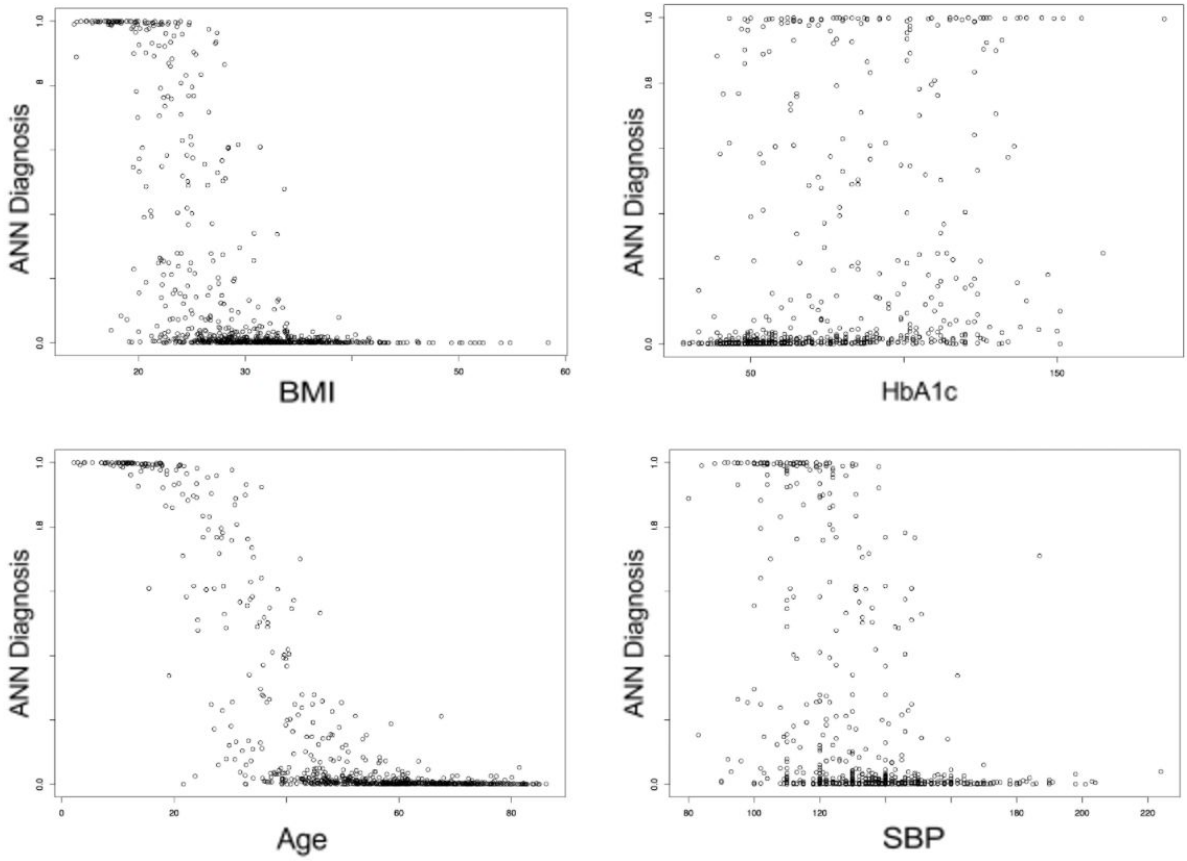

Figure 3 illustrates the association between examined patient characteristics and ANN model predicted diagnosis. BMI (top left), HbA1c (top right), Age (bottom left) and SBP (bottom right) are illustrated.

Figure 3 illustrates the relationship between the ANN algorithm developed and the basic factor weighting and associations made by the model per factor. As BMI increased from 20 to 40, the ANN predicted diagnosis gradually changed from Type 1 to Type 2, therefore concluding that an increase in BMI resulted in a Type 2 diagnosis. Of the factors examined, HbA1c illustrated the least pattern orientated ANN diagnosis, but it could be determined that as HbA1c values increased, there was an increased likelihood of Type 1 diagnosis. Age illustrated a direct correlation with ANN diagnosis. Below 20 years, ANN diagnosis correlated with Type 1 diagnosis, and as age increased from 20 to 60 years, ANN diagnosis changed from Type 1 to Type 2. Age 60 years and above illustrated a high correlation with Type 2 diagnosis. Finally, SBP showed correlation between an increase in Systolic BP and Type 2 diagnosis.

Results collected from the ANN model correlate with the basic Type 1 and Type 2 patient characteristics noted in Table 1, illustrating the ANN model has classified these results from the dataset and utilised the factors to obtain the observable diagnosis results. Table 1 demonstrates the mean age of Type 1 patients as 29 years, and Type 2 patients was 59 years. This correlates directly as average points on the Figure 3 illustration. Similarly BMI and mean Systolic SBP demonstrate an increase in Type 2 patients compared to Type 1, which again has been plotted in the ANN model diagnosis. Finally, mean HbA1c values of the patient Type 2 cohort illustrate a decreased value of 65, over Type 1 patients with a mean HbA1c value of 86. This further correlates with the model structure developed by the ANN model.

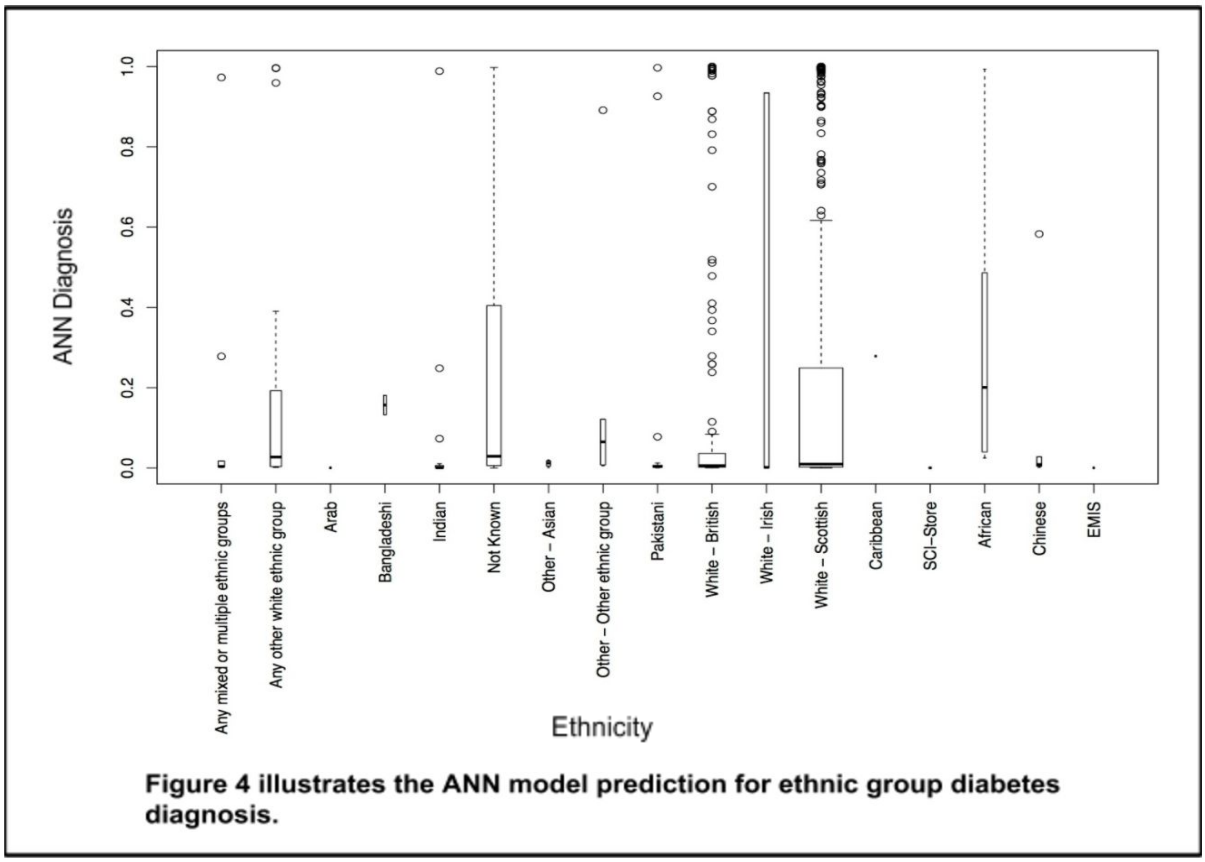

Figure 4 illustrates the ANN model prediction for ethnic group diabetes diagnosis.

Figure 4 illustrates the main ethnic groups present within the full dataset cohort, and the corresponding ANN model predicted diagnosis association. From the graph it is clear that African, ‘other ethnic groups’ and ‘other white ethnic groups’ are the 3 main ethnic categories with a mean diagnosis furthest from Type 2, therefore more likely to be Type 1 Diabetic than the other ethnic groups. Indian, Chinese and Pakistani all correlate with Type 2 diagnosis as mean ANN prediction is the lowest of all ethnic groups.

Again, information process by the ANN model can be directly correlated with the main ethnic groups of the dataset cohort. Notably, African patients have a 10.1% likelihood of Type 1 diagnosis, correlating directly with the ANN model prediction of Type 1. Furthermore, Indian and

Pakistani patients are most likely to have Type 2 diabetes, with only a 2.8% and 3.1% likelihood of Type 1 diabetes.

DISCUSSION

Results Analysis

Prior to analysis of the ANN and logistic regression model results, it was crucial to examine the cohort of patients within the general dataset, hence the statistical summary featuring the essential patient parameters. Examining the initial cohort allowed a direct comparison when analysing the functioning method of the ANN model. For example, when analysing the plot of BMI vs ANN diagnosis, the ANN model results illustrated that as BMI increased, the patient was more likely to be predicted Type 2. This demonstrated that the ANN was efficiently selecting information from the training set and correlating an increase in BMI with an increased likelihood of Type 2 diabetes. Demonstrating that the ANN model and current patient diagnosis from the dataset showed similar diagnostic patterns (Table 1 and Figure 4)illustrates the ability of the advanced neural network to identify important criteria and correlate with the correct diagnosis. Had the ANN disagreed with the trends present within the data, the model would not have been considered efficient at utilising the training data to weight the factors and therefore would have required further tuning.

The results collected illustrated a positive development in the advanced networking model in comparison to the classical logistic regression and the current physician approach. By comparing the accuracy, specificity and sensitivity of the competing models, it was clear the ANN was most efficient on every stage of analysis, with the physician diagnosis concluded as the least reliable. Both computerised models also outperformed the physician diagnosis when relative performance was examined, with the ANN and logistic regression illustrating a greater true positive and lower false positive rate than that of the physician. This demonstrates that the models are more efficient at correctly diagnosis the patients than the physician alone.

When comparing the advanced and classical models, the AUC score was identical at 0.95, with minor improvements of the ANN model compared to the logistic regression for accuracy, sensitivity and specificity. Although only minor improvements over the classical approach, the observed results are considered to be successful as the aim was to develop an advanced ANN that would compete to the same level as current models available, and the model developed has showed areas of superiority of the classical approach. Upon further model adaption the advancement of the ANN over the logistic regression may increase, however as a starting point for the model development the results collected are considered highly successful.

Model Limitations

Although considered the most efficient model examined, the advanced ANN faced some limitations when in development and therefore there are a number of areas of potential improvement.

When data was initially collected from SCI-Diabetes, a number of patients were presented with incomplete datasets. For the analysis undertaken, any patient with an incomplete dataset was excluded from the examinable cohort. To determine if this data exclusion was significant on the results collected, it would be necessary to repeat the analysis including previously excluded patients and undertaking data imputation to complete any missing data. This would also allow an examination into the cohort of individuals with incomplete datasets, and would determine if there was patterns present within those that had missing data.

The model could potentially be further improved by altering the overall ANN structure. The current analysis model was formed of 2 input levels at 16 units each and an output function, with an Epoch of 50 cycles. By increasing the number of levels, the number of units per level or altering the number of cycles the neural network would complete, the model could potentially increase in accuracy. The current AUC value of 0.95 is considered excellent for prediction however, and altering the model would only slightly improve the AUC to a maximum of 0.97 due to the limitations of quality of the SCI-diabetes dataset.

Furthermore, a number of classification thresholds and time limitations of data included in the analysis may have limited the model. It would be necessary to repeat the initial data analysis altering these values to determine the most efficient model that could be developed. For the undertaken analysis, the window of data collection was 3 months prior and following diagnosis date. By decreasing this window to 1 month, accuracy of the model may increase as the data selected may correlate more directly with initial levels upon diagnosis. Additionally, a classification threshold of 0.1 was determined, and individuals with an ANN predicted value of 0.1 or greater were classified as Type 1. By altering this threshold, the number of true positive and false positive Type 1 patients diagnosis would be altered, affecting the ROC curve and AUC value.

Finally, the ANN model could be improved by increasing the sample size to include SCI-Diabetes data from a number of regions, for example the addition of NHS Ayrshire and Arran patients. This would also be useful to compare the results between regions of Scotland, and assess if factors are weighted differently in different regions of the country.

Future Development

Data Quality

A major issue highlighted by the results collected from the ANN model was the issue of data quality of current patient diagnosis. With the ANN and current diagnosis in concordance for only 93% of the dataset, 3,500 patients were highlighted as having an inconsistent diagnosis

between current diagnosis and ANN diagnosis prediction. A beneficial use for the patients included in the data analysis undertaken would be to assess individuals that have a current diagnosis that differs from the ANN, logistic regression and physician diagnosis. There should be major concern over the accuracy of the current patient diagnosis should all 3 models dispute it, therefore this method would highlight patients with the greatest likelihood that current diagnosis is incorrect and require re-assessment.

Clinically, this type of analysis could be implemented on a large scale upon further ANN development. In a clinical setting, patients would only require the one ANN model to produce a diagnostic prediction that could be considered a reliable indicator of the correct diagnosis based on the criteria examined. The model could then be undertaken by physicians on their patient dataset, and individuals highlighted if ANN prediction differs from the current diagnosis. This would then allow the physician to re-assess all of the highlighted patient's available data, looking outwith the examinable criteria held by the ANN. This would therefore allow a fully informed decision to be reached by the physician as to the accuracy of the current diagnosis. Such a system would be highly beneficial if utilised in databases such as SCI-diabetes, by highlighting the individuals with the highest likelihood of misdiagnosis from a large database as all patient information would be easily available to the physician.

The development of a computerised diagnostic analysis model would be an important advancement in the efficient diagnosis of diabetic patients. Currently, patients are diagnosed and treated as determined by their physician alone, and are re-assessed only when issues are raised or treatment deemed unacceptable. This model would create an easy-to-use system that would allow instant examination over a database of patients, creating a list of those patients most likely misdiagnosed, and therefore creating an easier way to identify and effectively treat patients in an efficient timeframe.

SOM and ANN combination

Preliminary studies by NHS Greater Glasgow and Clyde Diabetologist physicians have shown an improvement on results collected by combining the use of unsupervised machine learning techniques with the ANN model. By initially undertaking an unsupervised approach to analysis the dataset then applying the results to an ANN, a list of predicted probability of misdiagnosis for each individual within the examined dataset can be accurately compiled.

The development of an SOM model has shown to be efficient, by highlighting unidentified associations within the data. By undertaking the SOM on the initial training set to identify the further associations, the ANN would then be able to accurately assess the information and create a more efficient model than when undertaken alone.This combination method could be considered highly successful clinically if developed to an efficient standard, by presenting the physician a highly accurate ranked list of patients which have a diagnosis that should be reassessed.

Although in first stages of development, the initial findings of the combination approach have been considered highly successful. By further developing the ANN model efficiency, the combination of approaches could offer a highly useful tool in the field of diabetes diagnosis.

CONCLUSION

The aim of the project was to develop an ANN diagnostic model for Type 1 and Type 2 diabetes that was as accurate as current models available. The results have shown the the unsupervised ANN developed was the most accurate model examined, with considerable efficiency over the physician diagnosis results. AUC values illustrated a model as accurate as the classical approach, with other examinable factors such as sensitivity and specificity illustrate a superior development. By further developing the model and considering a combination approach of unsupervised and supervised machine learning methods to increase accuracy, a system to rank patients based on misdiagnosis probability is an easily achievable goal.

The potential clinical benefit of the implementation of an advanced ANN model throughout networks such as SCI-Diabetes would be highly beneficial in increasing positive diagnosis in patients. Utilisation of this system would be a major step toward integrating and positively utilising Artificial Intelligence within the healthcare system, once fully developed to maximum capability.

ACKNOWLEDGEMENT

I would like to thank my supervisor Dr Christopher Sainsbury for his continuous support throughout my project and for introducing me to a field of study with such significance to the current medicinal market. His enthusiasm for the area of Machine Learning made my introduction to the topic a wonderful experience. I would also like to thank Dr Gregory Jones for including me into his research team, and discussing the potential opportunities in the field of AI and medicine. Finally, I would like to thank the Physicians whom carried out our diagnostic survey, as the project would not have been possible without them.

REFERENCES

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452–459. doi: 10.1038/nature14541

Hodeghatta, U. R., and Nayak, U. (2017). Unsupervised Machine Learning. In *Business Analytics Using R - A Practical Approach* (pp. 161–186). Apress.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69.

Maglogiannis, I, G. (2007) Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, *Information Retrieval and Pervasive Technologies.* IOS Press.

McCarthy, J. (2001). What is AI? / Basic Questions. Retrieved July 24, 2017, from http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html

Oakden-Rayner, L., Carneiro, G., Bessen, T., Nascimento, J. C., Bradley, A. P., and Palmer, L. J. (2017). Precision Radiology: Predicting longevity using feature engineering and deep learning methods in a radiomics framework. *Scientific Reports*, *7*(1), 1648.

Xiang, Z., Schwartz, Z., Gerdes, J. H., and Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management*, *44*, 120–130.

R PACKAGES USED:

Dowle M and Srinivasan A (2017). _data.table: Extension of `data.frame`_. R package version 1.10.4, <URL: https://CRAN.R-project.org/package=data.table>.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J and Müller M (2011). “pROC: an open-source package for R and S+ to analyze and compare ROC curves.” _BMC Bioinformatics_, *12*, pp. 77.

RStudio Team (2016). _RStudio: Integrated Development Environment for R_. RStudio, Inc., Boston, MA. <URL: http://www.rstudio.com/>.

R Core Team (2017). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <URL: https://www.R-project.org/>.

Sing T, Sander O, Beerenwinkel N and Lengauer T (2005). “ROCR: visualizing classifier

performance in R.” _Bioinformatics_, *21*(20), pp. 7881. <URL: http://rocr.bioinf.mpi-sb.mpg.de>.

Wing MKCfJ, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, Team tRC, Benesty M, Lescarbeau R, Ziem A, Scrucca L, Tang Y, Candan C and Hunt. T (2017)_caret: Classification and Regression Training_. R package version 6.0-76, <URL: https://CRAN.R-project.org/package=caret>.