Healthcare Workflows and Machine Learning

Luke Edgecombe
Robotics with Industrial Application
Heriot-Watt University
Edinburgh, Scotland
Email: le4005@hw.ac.uk

Chinyere Ihuoma Uwa
Applied Cybersecurity
Heriot-Watt University
Edinburgh, Scotland
Email: ciu4000@hw.ac.uk

Jahnavi Makaraju
Applied Cybersecurity
Heriot-Watt University
Edinburgh, Scotland
Email: jm4042@hw.ac.uk

Sarjjana Venkataramana
Applied Cybersecurity
Heriot-Watt University
Edinburgh, Scotland
Email: sv4016@hw.ac.uk

***Abstract*—This project explores using Machine Learning (ML) on three different medical datasets: brain MRIs to find tumors, a table stroke prediction data, and a combined dataset for predicting several diseases at once. We use specialized Convolutional Neural Networks (CNNs) for the images, and simpler classical models like Logistic Regression, Random Forest, K-Nearest Neighbors (KNN) for the clinical data tables. We run experiments on each set to see how things like data imbalance, feature representation, and model choice affect the outcome. The results show that different ML techniques work best with different types of data, and stress how important proper data cleanup, balancing the classes, and picking the right features are for making reliable medical predictions.**

***Keywords*—Medicine, Machine learning, Classification, Optimisation, Image analysis**

I. Introduction

In this report, a selection of differing machine learning (ML) techniques are explored. Each technique was chosen based on its area of application. The subject area of medicine was chosen for exploration, and a selection of datasets are explored with different features and formats: (i) a collection of brain magnetic resonance imaging (MRI) images that exhibit three cancer-based pathologies; (ii) a tabular dataset containing patient demographic and healthcare-related information related to stroke pathology; (iii) ?.

Stroke represents a significant global health challenge; it remains one of the most common causes of death and disability [1]. The application of ML to stroke datasets has become increasingly prevalent as it enables clinicians to identify risks earlier and make more informed treatment decisions [2]. The objective was to evaluate various ML models to determine their classification performance and assess how imbalanced data impacts prediction accuracy.

The accumulation of valuable medical data has become increasingly common, aiding early detection of health issues and effective treatment [3]. Hospitals are gathering large volumes of structured data, including patients' demographic information, hospital stay details, billing records, and basic health indicators. Such data can also be exploited by machine-learning algorithms to identify patterns that clinicians might not otherwise notice [4]. By analysing these patterns, ML supports clinical decision-making, identifies potential risks, and expedites the diagnostic process [5]. This project evaluates the performance of machine-learning models in predicting medical conditions using structured healthcare data [6].

To approach the ML problems, an understanding of previous literature on the topic is required for an effective workflow. Academic papers were reviewed to gain an effective skill set and background knowledge required to begin experimentation.

II. Literature

ML as a concept is the basis of the field of artificial intelligence (AI); the primary purpose of ML is the generation of algorithms that are able to learn. In this regard, research has focused on the optimisation of these systems by generation of new techniques and identifying the most optimal combinations of pre-existing software/hardware [7].

ML is a popular tool in stroke research and its used to help spot problems earlier and assess patients more accurately. Large-scale research clearly shows that age, blood pressure, heart conditions, and blood pressure all play a role in how stroke affects a patient, hence the use of predictive modelling is needed to identify issues and intervene as early as possible [1]. Traditional statistics methods, like logistic regression are popular in healthcare because they're easy to interpret and identify factors that contribute to a patient's risk.

A. Logistic Regression

According to Aboong [8], logistic regression effectively identifies the primary stroke risk factors, demonstrating its utility for analysing organised medical datasets. Heo et al. [2] developed an ML model for predicting stroke outcomes, demonstrating that ML methods outperform conventional scoring tools. A significant challenge, however, is that stroke datasets typically exhibit severe class imbalance, which can adversely affect model performance. Employing techniques such as Synthetic Minority Over-sampling Technique (SMOTE) to balance the data has been shown to improve the model's ability to detect rare conditions in medical datasets [9].

The MRI image set relies on the classification ML technique; the specifics of this are introduced here.

B. Image Classifiers

This selection was reinforced by past research on brain cancer MRIs [10]. In particular, a convolutional neural network (CNN) was deployed. This ML technique has been shown to be highly effective in image classification [11]. A review into CNNs describes the typical architecture and construction techniques [12]. As the task differs, the design of the CNN

can change, although the general format remains essentially the same. This typical structure, shown in Fig. 1, is comprised of ordered convolutional and pooling layers that come together to make a feature extractor. The inputted data is converted into a feature representation. In combination with this, fully-connected neural layers are integrated with activation functions to perform the desired ML operation.

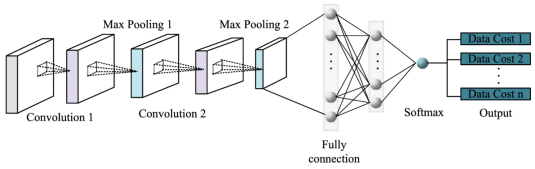


Fig. 1: Convolutional Neural Network Architecture [12]

Prior to data being used in ML, in order to increase suitability and feasibility of the model training, it should receive some sort of pre-processing. In order to effectively train a model, the data should be affected in a way that benefits the robustness and accuracy of the training process [13].

1) Pre-Processing: In image classification, effective pre-processing has been shown to increase the rate of identification [14]. For very large datasets, where manual review is not possible, automatic pre-processing is a must. Studies show that image cropping and reshaping, image resizing, and background noise removal are beneficial to the training process.

III. Data Analysis and Exploration

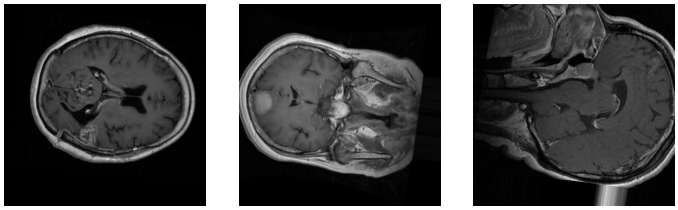
Based on the literature discussed previously, the data used in the following scenarios has undergone careful analysis in order to understand and prepare for upcoming ML. The MRI classifying dataset was sourced from a comprehensive collection of MRI images collected from a series of hospitals in Bangladesh [15]. The data holds great value and significant effort was made to collect and label high quality images.

A. Brain Cancer MRI's

A sample of the three classes of data present is shown in Fig. 2, pictures (a to c). The proprietor has already uniformly resized the images to equal dimensions. The ML for this set uses the hold-out method; 70% of the data is used for training, with the remaining 30% divided equally between validation and testing. To avoid bias or overfitting the data is shuffled using an scikit-learn cross-validation function that returns stratified randomised folds [16]. For the three sets some further ML specific pre-processing steps are executed, with training having more:

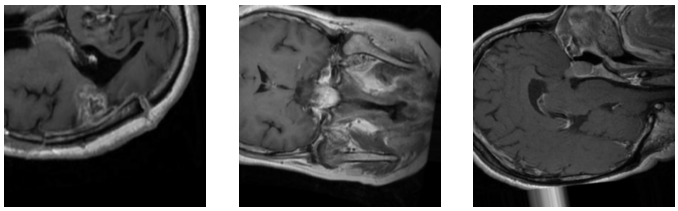
- Random resized crop to 224×224 pixels
- Random horizontal flip with 50% probability
- Conversion to 3-channel greyscale
- Conversion to tensor format
- Normalisation to floating-point values in range [0,1]
- Standardisation using ImageNet mean and standard deviation

Additionally, the stroke dataset presented some inherent issues.



(a) Glioma (b) Meningioma (c) Tumor

Fig. 2: Sample images from MRI scans



(a) Glioma (b) Meningioma (c) Tumor

Fig. 3: Preprocessed sample images from MRI scans

B. Stroke patient data

The stroke dataset used contains 5,110 patient records with demographic and clinical attributes such as age, heart disease, BMI, hypertension and average blood sugar level. The main variable (stroke) is binary. A small amount of missing BMI data was filled in by using the mean value.

A major challenge with this dataset is its severe imbalance, with only approximately 3.4% of cases representing actual stroke occurrences. The data exploration indicated a higher probability of stroke among individuals within the age range of 50–80 years, as shown in Fig. 4, and those with elevated glucose levels. When correlation analysis was carried out, it showed that age, hypertension and heart disease had subtle but meaningful relationship with stroke occurrences.

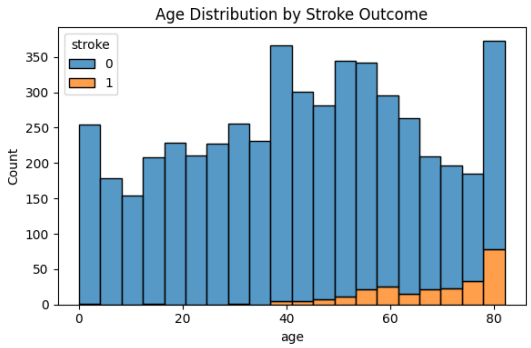


Fig. 4: Age Distribution by Stroke Outcome

To get the data ready for training, word-based features were converted to 0/1 columns so the models could process them. All numerical values were also rescaled to a similar range to ensure no feature dominated the other while the model was trained. To deal with the class imbalance problem, SMOTE was used to generate synthetic stroke cases, which helped

in detecting rare instances of stroke. The preprocessing and exploration guided the modelling choices in later stages.

After the initial data-processing step has been completed for all the datasets, the ML algorithms can now be established.

C. Combined Healthcare Dataset

This study also considers a combined healthcare dataset assembled for multiclass disease prediction using non-clinical attributes and symptom indicators. We treat the administrative/demographic records and symptom-based records jointly as a single, combined dataset for analysis and modelling.

Within the combined dataset, the symptom-based component is concerned with general disease prediction using data on symptoms. It is a large array of symptom features, whereas the last column is the diagnosed disease. The problem is represented as a multiclass classification task in which the model will be required to estimate one category of disease through a combination of symptoms. The dataset is used to train machine-learning algorithms like the Random Forest classifier to learn the relationship between patterns of symptoms and the occurrence of certain diseases and to assess the ability of symptoms alone to aid in the accurate prediction of a disease [5], [6].

1) Healthcare multiclass dataset: The healthcare multiclass portion includes 55,500 patient records, each labelled with one of six medical conditions: Arthritis, Asthma, Cancer, Obesity, Hypertension, or Diabetes. Each condition accounts for roughly 16–17% of the entire number of samples and, therefore, the dataset is perfectly balanced. To explore the range and preparedness of these demographic and administrative properties, we explored variables including age, sex, blood type, insurance provider, billing amount, and length of stay to assess variability. No empty or duplicate entries were noticed and verification that the dataset is clean and complete was done. The pie chart in Fig. 5 shows the even spread among the six medical conditions, to help make sure that when the model learns from the data, it should not favour any one class. Now this dataset is non-clinical, but it does provide a good foundation to explore how well the demographic and administrative attributes support multiclass disease prediction [3], [4].

2) Symptom-based disease prediction: The symptom disease dataset is a pool of patient records of symptoms, and the last column is the disease that the patient has. The data features several binary symptom attributes that represent the absence or presence of a given symptom and one target attribute that represents the disease that was diagnosed. It was verified that all the entries are complete and there are no missed or repeated values, which proves the data is clean and can be used for training. To get familiar with the data structure, the distribution of diseases over the dataset was analysed. All the disease categories are adequately represented and, therefore, the model will get balanced samples during training and will not be biased in any way to one type of classification. The variability of the symptom features is also good and this enables the classifier to learn various symptom patterns of varying diseases.

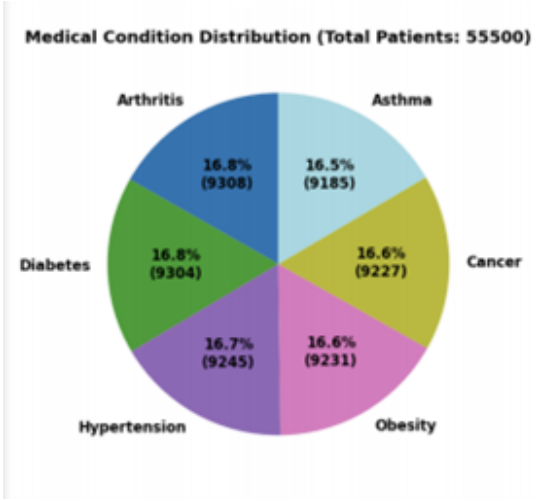


Fig. 5: Patient distribution pie chart

Even though there are no clinical measurements or numerical health indicators available in the dataset, the large variety of symptoms gives a solid foundation on multiclass disease prediction [5]. This makes the dataset appropriate for training machine-learning models, e.g., the Random Forest classifier [6].

IV. Experimental Setup

As the content of the datasets varies in format and shape, the experimental setup for each varies too.

A. Image classification training and evaluation

The MRI dataset was trained using a convolutional neural network (CNN) implemented as a customisable PyTorch model called `MRIClassifier`, based on the architecture shown in Fig.1. The architecture consists of the following components:

1) **Initialisation and Configuration:**
   - The model accepts two configurable parameters: `num_classes` (default 3) to specify the number of output classes, and `act_func` (default `nn.ReLU`) to define the activation function used throughout the network.
   - An internal activation factory function is implemented to instantiate the specified activation function, automatically handling the `inplace` parameter when supported by the activation function.
2) **Feature Extraction Blocks:**
   The feature extraction component comprises four convolutional blocks arranged sequentially, each following the pattern: convolution $\rightarrow$ batch normalisation $\rightarrow$ activation $\rightarrow$ max pooling.
   - *Block 1:* Processes the input 3-channel image (RGB or similar) using a `Conv2d` layer with 32 output channels, 3×3 kernel size, and padding of 1 to preserve spatial dimensions. This is followed by `BatchNorm2d` for normalisation, the configured

activation function, and `MaxPool2d` with stride 2 to downsample by half.

- *Block 2:* Increases feature depth from 32 to 64 channels using a `Conv2d` layer with 3×3 kernel and padding 1, followed by batch normalisation, activation, and max pooling to further reduce spatial dimensions.
- *Block 3:* Expands feature maps from 64 to 128 channels through convolution with 3×3 kernel and padding 1, applying batch normalisation, activation, and max pooling.
- *Block 4:* The final convolutional block increases depth from 128 to 256 channels using a `Conv2d` layer with 3×3 kernel and padding 1, followed by batch normalisation, activation, and max pooling.

3) **Adaptive Global Pooling:**
   - An `AdaptiveAvgPool2d` layer reduces the spatial dimensions of the feature maps to 1×1, regardless of input size. This ensures a fixed-size feature vector of 256 elements enters the classifier, making the network adaptable to varying input dimensions.
4) **Classification Head:**
   - A dropout layer with probability 0.5 is applied for regularisation to prevent overfitting during training.
   - A fully connected linear layer (`nn.Linear`) maps the 256-dimensional feature vector to `num_classes` output logits (3 by default), producing raw class scores.
5) **Forward Pass Pipeline:**
   The forward propagation through the network follows this sequence:
   a) Input images pass through the four convolutional feature extraction blocks.
   b) The resulting feature maps are globally pooled using adaptive average pooling to 1×1 spatial dimensions.
   c) The pooled features are flattened into a 1D vector of 256 elements.
   d) The flattened vector passes through the classifier (dropout followed by linear layer) to produce class logits.
   e) The output logits are returned for subsequent loss calculation or prediction.

This architecture progressively increases feature depth (3→32→64→128→256 channels) while reducing spatial dimensions through max pooling, allowing the network to learn hierarchical representations from low-level edges to high-level semantic features for MRI classification.

The design of the training loop is configured to utilise the defined model. The path of the data is visualised and displayed in simple flow charts for training and testing, shown in Appendix Fig. 11 and Fig. 12. Additionally, during training the performance of the model over each epoch was recorded. To provide extra insight into the training and broaden the scope, the model was run with varying hidden layer activation functions.

For the stroke dataset, after the data was cleaned, three primary ML models were used.

B. Stroke analysis training and evaluation

The selected models were:

- K-Nearest Neighbours (KNN)
- Logistic Regression
- Random Forest Classifier

The data was split into two parts for training and testing; 70% was used for training the models and the remaining 30% was set aside to evaluate performance. The models were evaluated using four main metrics: accuracy, precision, recall, and F1-score. After applying SMOTE, the models were retrained. To ensure a balanced evaluation of the stroke data and to compare with the classical models, a multi-layer perceptron (MLP) was used to train the balanced data.

V. Results

For the MRI set, the final results of the performance of the model over the testing set can be visualised in the form a confusion matrix, shown in Fig. 6. The classification report summarises the model's quantitative performance (see Table I).

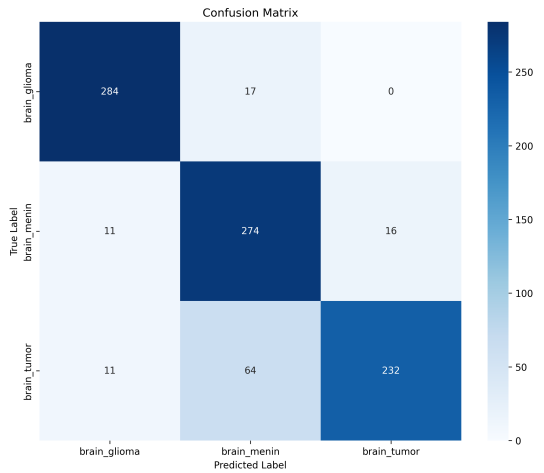


Fig. 6: Confusion matrix for brain cancer set (ReLU)

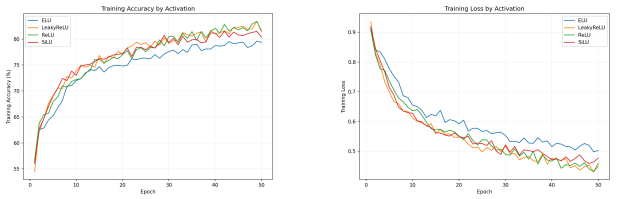
TABLE I: Classification report for CNN brain tumour prediction (ReLU)

Class	Precision	Recall	F1-Score	Support
Brain Glioma	0.93	0.94	0.94	301
Brain Meningioma	0.77	0.91	0.84	301
Brain Tumour	0.94	0.76	0.84	307
Accuracy	0.87	0.87	0.87	0.87
Macro Avg	0.88	0.87	0.87	909
Weighted Avg	0.88	0.87	0.87	909

The result of the differing acitvation function performances are displayed in a line chart showing the running accuracy and loss of the training set and a table of final epoch performance. Shown in, Fig. 7 (a) and (b), and table II.

TABLE II: Final epoch comparison of activation functions for CNN MRI model

Activation	Train Loss	Val Loss	Train Acc%	Val Acc%
LReLU	0.4499	0.3561	81.67	86.47
SiLU	0.4770	0.3747	80.37	85.48
ReLU	0.4588	0.4156	81.48	84.71
ELU	0.5023	0.7485	79.38	63.70



(a) Combined training accuracies for MRI set (b) Combined training losses for MRI set

Fig. 7: MRI set training performance

For the stroke prediction set, the following confusion matrix and classification report summarise the model performance on the test data.

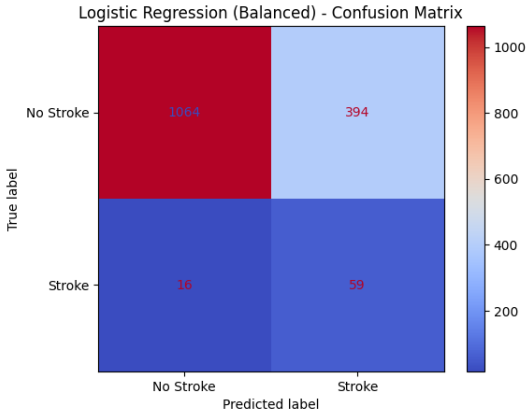


Fig. 8: Confusion matrix for logistic regression (SMOTE-balanced dataset)

TABLE III: Classification report for stroke prediction

Class	Precision	Recall	F1-Score	Support
No Stroke	0.99	0.73	0.84	1458
Stroke	0.13	0.79	0.22	75
Accuracy	0.73	–	–	1533
Macro Avg	0.56	0.76	0.53	1533
Weighted Avg	0.94	0.73	0.81	1533

A. Combined dataset results

The Random Forest classifier with only demographic and administrative features in the first model performed the worst. The accuracy was 0.16, and the precision, recall, and F1-scores were similar to 0.16–0.17. The output (i.e., predicting Obesity with probability 0.3835) shows that these non-clinical features provide little partition between categories. By contrast, the symptom-based model did a very good job. The accuracy with precision, recall, and F1-score within all diseases (around 97%) is demonstrated below.

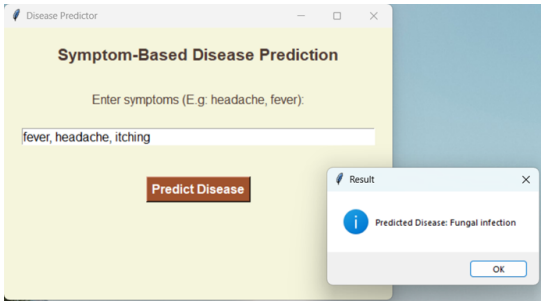


Fig. 9: Disease prediction

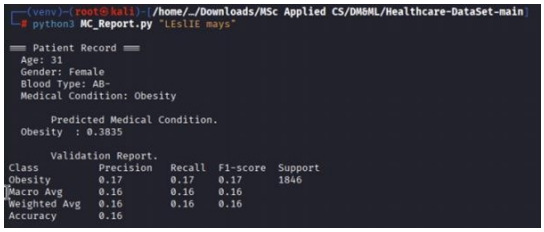


Fig. 10: Validation report print-out

VI. Discussion

The stroke prediction set revealed important insights into model performance. When the models were applied to the unbalanced data, the results showed high accuracy, but barely predicted any actual strokes (low recall).Which means they usually just guessed “no stroke” every time. Logistic Regression performed best overall as shown in Fig. 8 and table III, it has a strong recall score of 0.79 for the stroke cases while maintaining a good accuracy. This improvement shows how important it is to deal with unbalanced data when making predictions. The neural network achieved a high accuracy of 92% but struggled to identify actual stroke cases (low recall), which suggests it had difficulty in learning the patterns of the minority class.

The CNN model demonstrated strong performance in classifying brain tumour types. As shown in the confusion matrix (Fig. 6), the model correctly identified the majority of cases for each class, with particularly high accuracy for Brain Glioma (284/301) and Brain Meningioma (274/301). Misclassifications were relatively low, though the model tended to confuse Brain Tumour cases with Brain Meningioma (64 instances). Consistent with the classification report, the overall accuracy was 87%. An activation ablation further indicated that LeakyReLU offered the best generalisation (final validation accuracy approximately 86.5%), followed by SiLU (approximately 85.5%) and ReLU (approximately 84.7%), whereas ELU underperformed (approximately 63.7% validation accuracy) with a notably higher validation loss. We therefore

adopted LeakyReLU in the final CNN, likely benefiting from its small negative slope which mitigates dying-ReLU effects without the instability observed with ELU.

a) Combined dataset: results and discussion: The experiments exemplify how significantly feature quality influences multiclass disease prediction. Corresponding results in [3], [4] indicate that classification using demographic attributes alone does not allow for sound multiclass classification. By contrast, the symptom-based model did a very good job. The model is good at combining different combinations of symptoms (fever, headache, itching) to the appropriate disease (Fungal infection), which has been previously shown in prior work demonstrating the success of Random Forests when diverse, clinically relevant features are present [5], [6]. Overall, these results support that Random Forests are stable between tasks, but high-quality clinical/symptom features are vital for multiclass medical prediction as a whole.

The results from our work on all three datasets show that model performance is influenced by the data quality, feature representation and class balance. Every dataset presented a distinct issue: classifying images, handling imbalanced stroke predictions or predicting multiple diseases using a mix of demographic and symptom data.

The CNN model for the MRI dataset we used performed well with 87% accuracy overall and high scores for two of the tumour types. Most of the errors occurred between tumours that look alike, which showed that it was hard to distinguish them based purely on images. The comparison of activation functions revealed that LeakyReLU was the top performer, which highlights that minor architectural choices can noticeably affect a model's ability to generalise when working with limited medical image data.

The stroke prediction results demonstrated the limitations of using classical models on unbalanced clinical data. On the original unbalanced dataset, the models exhibited high accuracy but struggled to identify actual stroke cases, predominantly predicting "no stroke". After applying SMOTE to balance the dataset, Logistic Regression achieved a strong recall of 0.79 for stroke cases with an acceptable accuracy of 0.73. This represents a favourable trade-off in healthcare, as missing a true stroke case is more detrimental than experiencing a minor reduction in overall accuracy. In contrast, the MLP neural network achieved high accuracy (92%) but it missed most of the actual stroke cases (low recall). This suggests how important it is to handle the class imbalances first when trying to model rare medical cases.

The combined healthcare dataset demonstrated that feature informativeness plays a central role in predicting multiple diseases. When a Random Forest was trained using only demographic and administrative details (such as age or billing information), it failed to distinguish between the six diseases. This confirms that these general attributes are too broad for specific clinical predictions [3], [4]. However, when the same model was trained using symptom-based features, it achieved nearly perfect results, demonstrating that the symptoms provided rich and unique information, consistent with findings from other studies [5], [6]. This contrast clearly highlights that even the best model will perform poorly if the input features don't have enough clinical relevance.

Furthermore, our experiments demonstrated:

- CNNs are highly effective when working with complex images.
- Simpler models such as Logistic Regression and Random Forest remain valuable for organised tabular data.
- Balancing techniques such as SMOTE are essential for fair evaluations in imbalanced clinical settings.
- The quality and clinical relevance of features (such as symptoms) are more important than data volume or model complexity.

VII. Conclusion

In this project, machine learning (ML) was applied to three distinct healthcare datasets: MRI images, tabular stroke patient records, and combined healthcare records to determine how model choices, data preparation, and feature design affect performance. The results demonstrated that no single model performs optimally for every scenario; rather, successful performance depends on the type and structure of the data being analysed.

The CNN model used on the MRI data, performed well in classifying tumours. This shows that convolutional architectures is well-suited for medical imaging. In stroke prediction, Logistic Regression combined with data balancing (SMOTE) offered the best performance trade-off between overall accuracy and detecting stroke cases. This shows that simpler, easy-to-understand models can be very effective if trained correctly. For the last dataset (combined healthcare data), the Random Forest model did poorly with the demographic information but performed well when trained on relevant symptom data. This showed that using relevant attributes are important for accurate prediction.

The main findings from this project were as follows:

- Model performance is influenced by data preparation methods and the relevance of chosen features.
- Handling class imbalances is essential when modelling rare medical cases to ensure fair evaluation and reliable minority-class detection.
- Model choice is guided by the data modality.
- Deep learning performs optimally with rich, complex data such as images, whilst simpler models remain effective for organised tabular data.

Finally, these findings support the growing trend of using machine learning (ML) to assist doctors make better decisions. It helps provide reliable risk predictions, assists with diagnosis and making better use of the patient data made available.

References

[1] V. L. Feigin, B. A. Stark, C. O. Johnson, G. A. Roth, C. Bisignano, G. G. Abady, M. Abbasifard, M. Abbasi-Kangevari, F. Abd-Allah, V. Abedi, A. Abualhasan, N. M. Abu-Rmeileh, A. I. Abushouk, O. M. Adebayo, G. Agarwal, P. Agasthi, B. O. Ahinkorah, S. Ahmad, S. Ahmadi, and Y. Ahmed Salih, "Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the global burden of disease study 2019," *The Lancet Neurology*, vol. 20, no. 10, pp. 795–820, Sep 2021. [Online]. Available: https://www.thelancet.com/journals/laneur/article/PIIS1474-4422(21)00252-0/fulltext

[2] J. Heo, J. G. Yoon, H. Park, Y. D. Kim, H. S. Nam, and J. H. Heo, "Machine learning–based model for prediction of outcomes in acute stroke," *Stroke*, vol. 50, no. 5, pp. 1263–1265, May 2019. [Online]. Available: https://www.ahajournals.org/doi/10.1161/STROKEAHA.118.024293

[3] A. H. Salman and W. Al-Jawher, "A new multi-class classification method based on machine learning to document classification." in *2023 16th International Conference on Developments in eSystems Engineering (DeSE)*, 2023, pp. 605–610.

[4] IEEE, *2020 International Conference on Information Science and Communications Technologies (ICISCT)*. Proceedings, Nov 2020.

[5] S. Rani, P. Kumari, and S. K. Singh, "Machine learning-based multiclass classification model for effective air quality prediction," in *2023 IEEE IAS Global Conference on Emerging Technologies*, 05 2023, pp. 1–7.

[6] N. Sourov, F. H. Chowdhury, A. Redowan, M. Tarifuzzaman, and D. M. Farid, "Ensemble machine learning for multi-class classification tasks," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 2023, pp. 1–6.

[7] R. Abdulkadirov, P. Lyakhov, and N. Nagornov, "Survey of optimization algorithms in modern neural networks," *Mathematics*, vol. 11, no. 11, 2023. [Online]. Available: https://www.mdpi.com/2227-7390/11/11/2466

[8] M. Aboonq, "Potential role of logistic regression analysis to identify significant risk factors associated with stroke," *Bulletin of Egyptian Society for Physiological Sciences*, vol. 44, no. 1, pp. 17–28, Jan 2024.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 16, pp. 321–357, Jun 2002.

[10] J. Kang, Z. Ullah, and J. Gwak, "Mri-based brain tumor classification using ensemble of deep features and machine learning classifiers," *Sensors*, vol. 21, no. 6, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/6/2222

[11] P. Wang, E. Fan, and P. Wang, "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning," *Pattern Recognition Letters*, vol. 141, pp. 61–67, jan 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865520302981

[12] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artificial Intelligence Review*, vol. 57, no. 4, Mar 2024. [Online]. Available: https://link.springer.com/article/10.1007/s10462-024-10721-6

[13] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, "The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis," *Computers in Biology and Medicine*, vol. 128, p. 104129, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482520304601

[14] P. M. Diop, J. Takamoto, Y. Nakamura, and M. Nakamura, "A machine learning approach to classification of okra," in *2020 35th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, 2020, pp. 254–257. [Online]. Available: https://ieeexplore-ieee-org.hwu-ezproxy.idm.oclc.org/document/9183312

[15] M. M. Rahman, "Brain cancer - mri dataset," *Mendeley Data*, vol. 1, Aug 2024. [Online]. Available: https://data.mendeley.com/datasets/mk56jw9rns/1

[16] scikit learn, "sklearn.model_selection.stratifiedshufflesplit." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html

Appendix A
Training and Testing Loop Diagrams

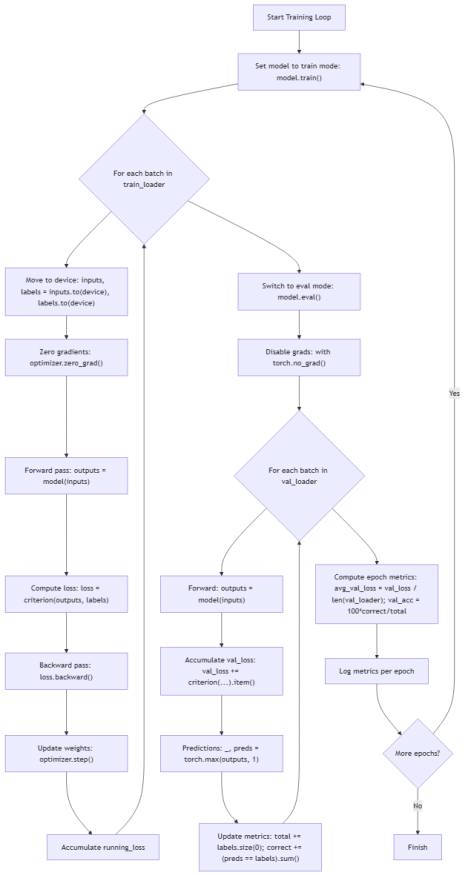


Fig. 11: Training Loop for MRI Classifier

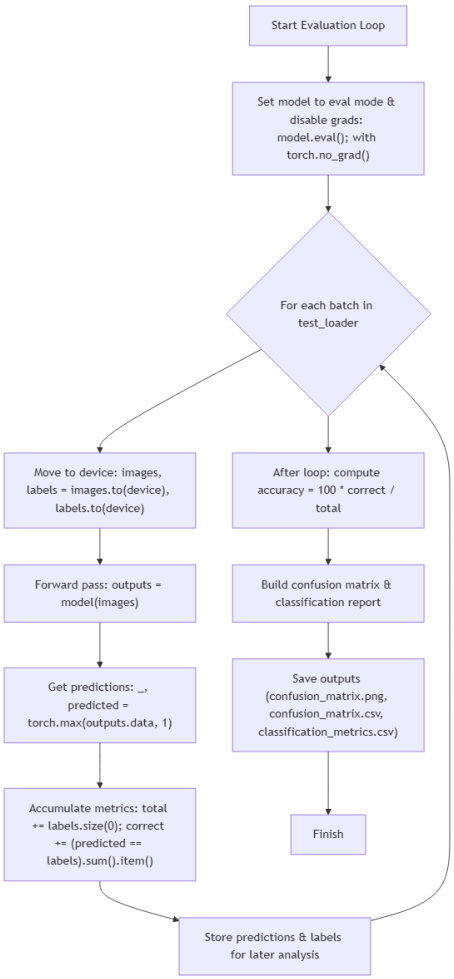


Fig. 12: Testing Loop for MRI Classifier