**Health Prediction Leverages Variety of Machine Learning Algorithms: A Comprehensive Review**

Kah Chun Wong1, a) and Syed Qamrun Nisa1, 2, b)

*1Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, Cyberjaya, 63100, Malaysia*

*2Centre for Image and Vision Computing, CoE for Artificial Intelligence, Multimedia University, Persiaran Multimedia, Cyberjaya, 63100, Malaysia*

*b) Corresponding author: qamrunnisa@mmu.edu.my*

*a) 1221304973@student.mmu.edu.my*

**Abstract.** This paper reviews the implementation and application of various machine learning algorithms and technique for health prediction. It focuses primarily on supervised learning algorithm such as K-Nearest Neighbour (KNN), Logistic Regression, Support Vector Machine (SVM), Convolutional Neural Networks (CNN), Decision Tree and other model, analysing their effectiveness and ability in disease prediction tasks. A set of performance matric including accuracy, precision, recall, and F1-score provide performed as a standard to evaluate and asses the compare model performance. Based on the reviewed articles and research, models that combining the Fuzzy Min-Max (FMM) and Random Forest (RF) techniques had demonstrated an outperformed predictive performance. These reviews concluded that a proper and effective data preprocessing, suitable model selection based on dataset are crucial for the purpose of developing trustable and accurate health prediction systems.

# **INTRODUCTION**

Rapid development of technology in this era has been embedded and integrated in many industrials such as finance, retail, agriculture, medical field where data mining has gained prominence. According to the information of World Health Organisation (WHO), data mining having potential for early disease detection and personalized treatment. Machine Learning (ML), a subfield of artificial intelligence (AI) further enhances it by analyse mined data to find out the market trend, discover the interesting pattern and hidden relationship from a huge amount of data (database), in order to interpret a meaningful data. [1].

One major change in medical epidemiology is the utilization of big datasets and ML models for the enhanced patient risk assessment, illness risk factor determination, and infection spread monitoring [2]. Different ML algorithms including KNN, Logistic Regression, Decision Trees, and Neural Networks provide a wide range of methodologies in identifying ways through which data mining and predictive modelling can be established. Notably, neural networks offer an advanced mechanism for medical decision support because, directly, they learn complex relationships between symptoms and diseases thus diminishing involving the expensive process of knowledge acquisition [3].

Moreover, integration of ML alongside real-time healthcare systems allows active and dynamic decision-making such as early warning of severe health states and resource allocation optimisation in hospitals. An example that we can use is the rate of patients getting hospitalized is something that can be predicted using predictive analytics so that they will be able to make decisions such as staffing and allocating resources better. In addition, remote patient monitoring and management of chronic diseases can be accomplished through wearable health devices and the Internet of Things (IOT) that create a constant flow of health data which could be analysed using ML methods. Not only does this combination of data mining, ML, and digital health solutions increase the efficiency of the healthcare provision process, but it also leads to better patient outcomes due to delivery of personalized and timely interventions.

In this paper, we aim to review the existing literature on health prediction utilizing different machine learning algorithms. We also intend to evaluate the performance of various machine learning algorithms and identify the most effective approach among them for accurate predictions.

# **BACKGROUND**

This section delves into the key techniques that applied in health prediction, offering a briefly description of each. It is structured into the following parts: 1. Machine Learning, 2. Supervised Learning, 3. Unsupervised Learning, and 4. Machine Learning Algorithms.

## **Machine Learning**

ML utilizes algorithms to build models that learn complex patterns from preprocesses data, ultimately generating predictive insights. Recent research highlights the value of these intelligent classifiers in medical and healthcare decision-making, including prognosis, diagnosis, and screening [3]. The ML process begins with clean data for training and testing. To address challenges like high dimensionality and noise, data cleaning and preprocessing are crucial steps to prevent skewed model results [2]. Following this preparation, the data is used to train and test models using both supervised and unsupervised learning techniques.

## **Type Of Machine Learning**

ML, as one part of the artificial intelligence (AI), having a powerful technique which enables the creations of computer systems that can learn from prior experiences without the need for programming or case studies. [4]. It plays a crucial role in terms of the classification and predictive analysis development by applying several techniques such as deep learning, evolutionary learning, reinforcement, and supervision. These insights are applied to quickly classify large amounts of data [5]. In this paper, we mainly divide the ML into two types: which are supervised learning and unsupervised learning

## **Supervised & Unsupervised Machine Learning**

Supervised learning is a technique that the model feed with a dataset that contains attributes and come along with the outcome or result accordingly, as known as labelled data. The model learns from the dataset during the model training. After trained, the model is ready to made prediction on new or unseen data based on what is learn from the labelled dataset. Normally, the common task for supervised learning algorithm is to perform for classification and regression tasks. Unsupervised learning in contrast feed with the dataset that without any prior knowledge of the correct and proper answers. However, an unsupervised leaning normally made prediction on the new and unseen data based on finding the trend, pattern, relationship, correlation and so on. This technique identifies the commonalities among input values and classifies them according to these commonalities. An unsupervised learning has a clustering technique.

## **Classification Of Machine Learning**

As mentioned before, classification is a technique that under the category of supervised learning. To perform predictions using ML techniques, a diverse array of algorithms has been widely implemented. In this section, several supervised learning algorithms be briefly explained, focusing on their operational principles in generating predictions.

## **Support Vector Machine (SVM)**

The biggest feature of SVM is the hyperplane that used to classifies the labelled training data point. The hyperplane act as a class separator and decision boundary, formed between two vectors from two distinct target class (e.g. “0” or “1”). SVM aim to seek the maximum margin (width of the hyperplane) from the closet vectors from each class. The larger margin of availability may enhance the model’s generalization capability, achieving reduced overfitting and improved functionality in unseen data. Classification for SVM can be made by determine which side of the hyperplane for the new data point belongs to. SVM can perfectly suited if the separation of classes has a clear margin [6].

## **K-Nearest Neighbour (KNN)**

KNN is a versatile algorithm capable for both classification (class labelling) and regression (predicting numerical value). KNN select the K closest data vectors from its surrounding data vectors. The proximity of these K neighbours is typically determined using the Euclidean distance formula, which quantifies the geometric distance between data points in the feature space. To perform classification tasks, KNN employs a majority voting principle. The new and unseen data is then be assigned to class label that is most frequent among its 'K' nearest neighbours. [2].

## **Logistic Regression (LR)**

Logistic regression is another powerful classification technique, fundamentally relying on using the probability estimation the occurrence of the event, whether that event occur or not [7]. A set of formulaic calculation involved to perform probability measure. Starts with receiving input features (Xi) and forming a linear combination (z = β₀ + β₁x₁ + β₂x₂ + ... + βₙxₙ) with their respective coefficients (βi). The sigmoid function is then applied into the formed linear combination, resulting a continuous probability score that between 0 and 1. Lastly, a separate rule often as a threshold with value 0.5 is employed used to transform the probability score into the binary class (class 1 or 0).

## **Decision Tree (DT)**

The image of algorithm behind DT just like its name, the features keep spanning like a tree structure to perform both classification and regression decisions. As a supervised learning, DT able to process both numerical and categorical data [8]. In DT Feature represents nodes, branches mean outcomes and leaves match the predicted class or value. Using criteria like variance reduction or Gini impurity, the algorithm iteratively divides the data according to features that best distinguish the target variable. In order to make predictions, the input features are used to navigate the tree until a leaf node is reached [9].

## **Convolutional Neural Network (CNN)**

Representing an instance of artificial neural network, Convolutional Neuron Network (CNN) is a network that can define patterns and keep forecasts according to the features. The CNN is characterized by complex segments that engage with one another such that to compute data. Convolutional layers, or hidden layers, involve the adoption of filters that are used to scan across the input data matrices so that patterns can be obtained and attribute correlations examined, such as between age, weight and exercise behaviours. The greater the development of the network, the more complex these filters be. Through the use of ReLU among other non-linear transformations, long-range relationships and input features are better harnessed; moreover, the pooling operation reduces dimensionality with a corresponding reduction in the amount of computation. Compressed and cast as unary representation the feature maps are integrated using the fully connected layer in order to estimate attribute classes. The end of the network’s output layer is for classification – SoftMax non-linearity is used for multi class case, sigmoid is used for binary class problem. In this way, CNN can exploit correlation structures and rank attributes in high dimensional data.

# **PERFORMANCE EVALUATION MATRIX**

Performance evaluation matrix is a common and standard matrix that widely been used to evaluate the performance of the model and accuracy of prediction result. Below section describe the most popular evaluation procedures and how they access the performance. The Table 1 provides a detailed explanation of how each metric evaluates errors, along with its corresponding formula.

# **RELATED WORK FOR HEALTH PREDICTION**

The studies paper included in this review were selected through a systematic search of online databases such as IEEE Xplore, Google Scholar and so on. We searched for relevant article using several keywords such as “machine learning”, “health prediction”, “data classification” and so on. We mainly focused on the article that followed typical pipeline and structure: discussing a specific disease or health issue, applying ML techniques, describing the data collection and preprocessing steps, training a proposed model, and finally comparing the results with other models.

|  |  |  |
| --- | --- | --- |
| **TABLE 1.** Common classification evaluation metrics | | |
| **Matrix** | **Explanation** | **Formula** |
| Accuracy | How accurate the overall prediction performance of model |  |
| Precision | From the predicted value as positive, how accurate is it. |  |
| Recall | How accurate the model to recognise the positive class |  |
| F1 | The combination of precision and recall, it assesses how well a model strikes a balance between accurately identifying positive events and avoiding incorrectly labelling negative cases as positive. |  |

Hybrid intelligent method combining Gaussian Mixture Model (GMM), principal component analysis (PCA), and Classification & Regression Tree (CART), and Fuzzy Min-Max (FMM) by [1] achieving extremely high accuracy (91-93%) on UCI datasets, outperforming PCA-SVM and PCA-KNN. [3] used CART, random forest, and FMM also obtained high accuracy up to 98.94% for breast cancer compared to standard FMM

[2] used CNN and KNN for chronic disease prediction, also reaching 96% accuracy on mixed dataset. [10] developed CNN-Based Multimodal Disease Risk Prediction (CNN-MDRP) and CNN-based Unimodal Disease Risk Prediction (CNN-UDRP) for cerebral infraction prediction. CNN-MDRP finally getting a higher accuracy with 94.80%. [11] also developed a CNN-based Multimodal Disease Risk Prediction (CNN-MDRP) model that integrates structured and unstructured medical data, alongside a CNN-based Unimodal Disease Risk Prediction (CNN-UDRP) model. Results showed that CNN-MDRP achieved a higher prediction accuracy with faster convergence compared to CNN-UDRP and other baseline methods Other authors also perform the comparison for CNN and KNN which goes to [12]. However, their study only state that the performance of CNN in terms of accuracy and time consuming is better than KNN

[13] want to identify the factors that influence the HRQoL (Health-Related Quality of Life) of the elderly with chronic diseases, and they found out that Stepwise logistic regression (SLR) is the most accurate (93%) in prediction chronic disease compare with SVM, DT, RF. [14] enhance the standard RF by combining SVM and long-short term memory (LSTM) features. Finally hitting a high accuracy with 97% in disease forecasting, which exceeded Naïve Bayes, KNN, and Linear Regression.

In a 2022 study, [15] turned their eyes to heart diseases, such as CAD. They compared the performance of SVM, ID3, C4.5, and Naïve Bayes, finding SVM is the most accurate algorithm with 92%. [6] used another method which is hyperparameter-tuned Logistic Regression and achieving 86% of accuracy for diabetes prediction, outperforming than KNN, SVM and RF. In the study of [16], they found out KNN also a good algorithm to perform heart disease prediction, but with the combination Logistic Regression, and the finally hit 87.5 of accuracy.

[9] compared 6 classification method (Logistic Regression, KNN, DT, SVM, NB) for liver disorders. Across all evaluated aspects, the Logistic Regression (LR) model demonstrates strong performance (75% accuracy, 91% precision, 83% F1-score). In addition, there are several studies also comparing the different classification method for their own dataset. In their respective studies, the different modules had the highest accuracy rates. For example, [4] claims that logistic regression (LR) is the best model to perform diabetes prediction with an accuracy of 96%. [7] finally tested that Ada Boost, SVM and LR is the best model to predict for breast cancer, diabetes, and heart disease with am accuracy of 98.57%, 85.71% as well as 87.10%, respectively. While [8] focus on the heart disease prediction. In the final result, random forest performed the best and produced a 99% of accuracy during the model training.

Across the studies, model such as CNN, SVM and ensemble methods like Random Forest tend to perform better in complex datasets due to their ability to capturing the non-linear relationship and pattern from the dataset. Accuracy often improved further when the model is combined (hybrid models), featured are well-engineered, or the hyperparameters are finely tuned. This suggest that model performance is highly dependent on the several factors such as quality of dataset, dataset & model handling, and also the specific health condition being predicted.

Table 2 below concluded a wide range of prediction accuracy and a comprehensive overview of article on health prediction by leveraging multiple machine learning model on various datasets, showcasing the feasibility of the models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TABLE 2**. Summary table on reviewed articles and research | | | | |
| **Authors** | **Data source** | **Data** | **Methodology** | **Accuracy** |
| M Nilashi et al. [1] | UCI ML Repository | Pima Indian Diabetes,  Wisconsin Breast Cancer,  StatLog,  Cleveland datasets | EM-PCA-Fuzzy Rule Based Technique | StatLog = 91.40% Cleveland = 92.80% Breast Cancer = 93.20% Diabetes = 92.90% |
| Manjeevan S et al. [3] | UCI ML Repository | Breast Cancer Wisconsin,  Pima Indians Diabetes,  Liver Disorders | FMM-CART-RF | Breast Cancer = 98.84% Pima Indians Diabetes = 78.39% Liver disorder = 95.01% |
| Rayan Alanazi et al. [2] | Real-life data collected obtained from various sources | Structured and unstructured data | CNN + KNN | 96.00% |
| Min Chen et al. [12] | Real-life hospital data from Wuhan, China | Structured and unstructured text data | CNN-UDRP algorithms CNN-MDRP algorithms | CNN-UDRP(T-data) = 94.20% CNN-MDRP (S&T-data) = 94.80 % |
| Lee S. K et al. [9] | Korea National Health and Nutrition Examination Survey (KNHANES) | Chronic diseases | SLR | 93.00% |
| K. Gaurav et al. [8] | Kaggle | Medical dataset | RF | 97.00% |
| Phasinam K. et al. [14] | UCI ML Repository | Heart Disease Cleveland | SVM | 92.00% |
| A.K.M S. et al. [4] | UCI ML Repository | Liver Disease Dataset | LR | 75.00% |
| Mujumdar & Vaidehi [5] | No | Diabetes Dataset | SVM, RF, DT, LR, KNN, NB, and so on | 96.00% (LR) |
| Dahiwade et al. [16] | UCI ML Repository | Disease Dataset | KNN, CNN | Accuracy of CNN > KNN |
| Yuvaraj & SriPreethaa [13] | National Institute of Diabetes | Diabetes Dataset | DT, NB, RF | 84.00% (RF) |
| Kohli & Arora [6] | UCI ML Repository | Wisconsin Breast Cancer,  Pima Diabates Disease,  Heart dataset | LR, DT, RF, SVM, AdaBoost | Breast cancer = 98.57% (AdaBoost)  Diabetes = 85.71% (SVM)  Heart Disease = 87.10% (LR) |
| Shah et al. [7] | UCI ML Repository | Cleveland Heart Disease Dataset | SVM, NB, DT, RF | 99.00% (RF) |

# **RESULT AND CONCLUSION**

This paper reviewed several studies on health prediction using different ML techniques. Various models were applied to different datasets, fielding a range of performance results. Among these Random Forest (RF) achieved the highest accuracy of 99% on heart disease dataset. The second model perform a highest of accuracy goes to combination of Classification and Regression Tree (CART), Random Forest (RF), and Fuzzy Min-Max (FMM) obtaining the accuracy with 98.84%, tested on the breast cancer Wisconsin dataset. Models using FMM also performed consistently well across other datasets, with an average accuracy of 92.57%, highlighting its effectiveness in health prediction tasks.

However the reviewed article employed different methodologies and data processing pipelines, some involved feature selection, others used data balancing techniques. This makes direct performance comparisons become challenging. To improve model transparency and usability for health professionals, future research should focus on developing standardize benchmarking frameworks, incorporating more diverse or realistic dataset, exploring explainable AI techniques. These steps are essential to ensure long-term accuracy, reliability, and usefulness for practical prediction systems for real-world healthcare applications.

# **REFERENCES**

1. M. Nilashi, O.B. Ibrahim, H. Ahmadi, and L. Shahmoradi, “An analytical method for diseases prediction using machine learning techniques,” Computers & Chemical Engineering **106**, 212–223 (2017).
2. Alanazi, “Identification and Prediction of Chronic Diseases Using Machine Learning Approach,” Journal of Healthcare Engineering **2022**, 1–9 (2022).
3. M. Seera, and C.P. Lim, “A hybrid intelligent system for medical data classification,” Expert Systems with Applications **41**(5), 2239–2249 (2014).
4. Z. Tasnim, and F.M.J.M. Shamrat, “A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms,” INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, (2020).
5. Mujumdar, and V. Vaidehi, “Diabetes Prediction using Machine Learning Algorithms,” Procedia Computer Science **165**, 292–299 (2019).
6. P.S. Kohli, and S. Arora, “Application of Machine Learning in Disease Prediction,” in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, (2018), pp. 1–4.
7. D. Shah, S. Patel, and S.K. Bharti, “Heart Disease Prediction using Machine Learning Techniques,” SN COMPUT. SCI. **1**(6), 345 (2020).
8. K. Gaurav, A. Kumar, P. Singh, A. Kumari, M. Kasar, and T. Suryawanshi, “Human Disease Prediction using Machine Learning Techniques and Real-life Parameters,” IJE **36**(6), 1092–1098 (2023).
9. Lee, Soo-Kyoung, et al. "Prediction model for health-related quality of life of elderly with chronic diseases using machine learning techniques." *Healthcare informatics research* 20.2 (2014): 125-134.
10. Singh, and R. Kumar, “Heart Disease Prediction Using Machine Learning Algorithms,” in *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, (IEEE, Gorakhpur, India, 2020), pp. 452–457.
11. S. Jadhav, R. Kasar, N. Lade, M. Patil, and S. Kolte, “Disease Prediction by Machine Learning from Healthcare Communities,” IJSRST, 29–35 (2019).
12. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning Over Big Data From Healthcare Communities,” IEEE Access **5**, 8869–8879 (2017).
13. N. Yuvaraj, and K.R. SriPreethaa, “Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster,” Cluster Comput **22**(S1), 1–9 (2019).
14. K. Phasinam, T. Mondal, D. Novaliendry, C.-H. Yang, C. Dutta, and M. Shabaz, “Analyzing the Performance of Machine Learning Techniques in Disease Prediction,” Journal of Food Quality **2022**, 1–9 (2022).
15. J.O. Healthcare Engineering, “Retracted: A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques,” Journal of Healthcare Engineering **2023**(1), 9872970 (2023).
16. D. Dahiwade, G. Patle, and E. Meshram, “Designing Disease Prediction Model Using Machine Learning Approach,” in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, (IEEE, Erode, India, 2019), pp. 1211–1215.