Predicting Heart Disease with Machine Learning Models

Brian Higgins[a],

[a] *MTU, Bishopstown, Cork, Ireland*

Abstract

Today, we are collecting huge amounts of medical data that we can combine with Machine Learning models to transform the healthcare industry. Machine Learning models can be used to predict various health conditions such as heart disease and cancer, reduce waiting times by assigning patients to the correct department and offer users health screening to indicate where lifestyle changes are needed. Heart Disease is the leading cause of death globally and early detection can play a curial role in its prevention. In this research, we took 4 heart disease datasets and used Logistic Regression, Random Forest and Gradient Boosting models to predict the likelihood of this condition.

Our results showed that the default settings of a Random Forest model performed the best with an accuracy of 84% on the validation data and 80% on the Test data. The Logistic Regression model performed with 82% accuracy on the validation data and 78% on the test data. With both of these models, we also performed a range of hyperparameters with grid search but with no improvements over the default models. Gradient Boosting scored in between with a validation accuracy of 82% on the validation data and a test accuracy of 78%. Using Grid with Gradient Boosting performed slightly worse with a validation accuracy of 82% and a test accuracy of 77%.

Research models need to have practical applications and so we designed a basic application where users can input their personal medical data which then makes a prediction of heart disease running on the best performing mode. This research was not just interested in finding the best performing model but also in delivering a practical real-world application.

Keywords
Machine Learning, Heart Disease, Logistic Regression, Random Forest, Gradient Boasting

1. Introduction

“Today approximately 30% of the world’s data is being generated by the health care industry” [1] which offers huge opportunities to have Machine Learning (ML) models gain insights and drive healthcare innovations in drug discovery, hospital management, virtual nursing and identifying high-risk patients all of which can be used to help with quicker diagnosis and reduce waiting times [2], [3].

Ireland has a modern health care system [4] with both public and private sectors, yet grapples with numerous challenges such as misdiagnosis [5] and growing numbers of people on waiting lists [6]. Long waiting times, patients spending days on trolls, crowded waiting areas and growing mistrust of doctors are all undermining the healthcare system. These problems negatively impact patient care but also create a two-tier system where more people pay for private health care, which then results in increasing similar issues in the private healthcare system [7].

Heart Disease is part of the Cardiovascular diseases and is the leading cause of death worldwide with an estimated 17.9 million lives lost each year [8]. This research aims to predict the likelihood of heart disease which can be used in both the prevention and diagnosis of this condition which can highlight the benefit of using ML and medical data. This research can also be used to reduce waiting times and other secondary benefits to health care systems.

1.1. Aims of the project

The primary aim of this project was to combine multiple heart disease datasets and apply various Machine Learning models to get the best-performing predictive model for heart disease. Another aspect was to research various Feature Selection Techniques and to see how they compare. Lastly, a goal was to create a practical application with the best-performing model to allow a user to enter their medical details and receive a prediction of their risk of having heart disease.

1.2. Description of the Data

The heart disease data is a combination of four datasets which give 920 observations:

- Cleveland Data – Cleveland Clinic Foundation – 303 observations
- Hungarian Data – Hungarian Institute of Cardiology, Budapest – 294 observations
- Switzerland Data – University Hospital, Zürich, Switzerland – 123 observations
- Va Data – Long Beach, VA Medical Centre – 200 observations

Each of these datasets has a processed version with 14 features and an unprocessed version with 76 features. Our models were run on the processed version looking to find the best-performing model which was then used in the practical application section. Examples of the processed features include age, sex, resting ECG, max heart rate and cholesterol. The result of these models were accurate to 80%. For the research, we looked at Feature Selection Techniques with the aim of looking to improve on the model's performance by exploring the larger 76 feature dataset version with the future aim of using these features to improve on the models performance.

1.3. Identify the Target

This research has used a binary target value of 0 for no heart disease and 1 for heart disease. 55% of the data is for heart disease and 45% of the data is for no heart disease. In this case, there is not a large imbalance for the target variable.

1.4. Previous Academic Work

The Cleveland dataset has been widely used in heart disease detection [9] and has been considered a benchmark dataset for heart disease detection. The UCI website [10] includes at least 40 research papers that have used at least one of the datasets we used in this research. These papers include a wide range of topics looking at compering Machine Learning Models, Neural Network Models and various other processing techniques

Several of the papers looked at different Machine Learning model comparisons [11], [12], [13] while others looked at Neural Network models [14], [15]. The goal of these research papers was to improve on the model's prediction and some of these research did not perform much better than this research paper. Others however did considerably better. A Support Vector Machine (SVM) model got an accuracy score of 89.57% [16], and another Deep Neural Network model performed with a 98.15% accuracy [17]. These papers offer interesting options for future research.

2. Research

This project aimed to run various machine learning models with Logistic Regression, Random Forest and Gradient Boosting to compare the performance of the models. They looked at Feature Selection methods to determine which features were the most important using the processed data of 14 features as a baseline to test each method. Using these methods the next step would be to use these on the unprocessed data of 76 features to try and improve the models performance. The larger unprocessed data needs a considerable amount of processing work so this has been noted for future work.

2.1. Feature Selection

Feature Selection Techniques can be an essential step when working through an ML pipeline. Feature Selection is the process of selecting the most appropriate features in your data that allows you to build the most appropriate ML models and obtain the most accurate performance results. Reducing the number of features has several benefits, including reducing the training time of models, simplifying models, reducing noise and minimizing overfitting. In this section, we will be looking at four techniques.

2.1.1 Heatmap with Correlation Matrix

This method uses a plot to show a colour range and a value for the correlations between all the pairs of features. It works well when there are a limited number of features but not so well when there are a large number of features. The use of a colour range and numeric values between -1 and +1 lets you see the highly correlated positive or negative features.

From our data, we have a heatmap in Figure 2.2.1 where we can see in blue positively correlated values which are coloured in green with the colour scale on the left. With the target feature, we have two moderately correlated values with Chest Pain Type (0.47) and Exercise-Induced Angina (0.43).

Negative Correlations are also important as seen with the black square, Max Heart Rate is negatively correlated (-0.35) with the target variable. A heatmap also allows us to see the correlation of features between each other, this way we can see features what might be worth combined. Age and Sex are an example where some research has showed a benefit in combining these features.

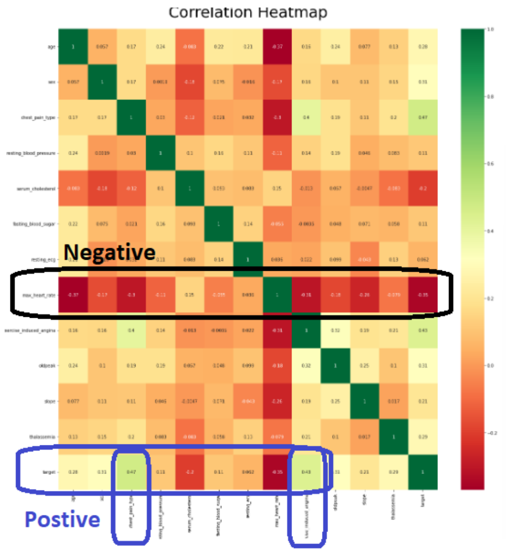

Figure 2.2.1: Heatmap

Random Forest Models	Accuracy
Model 1: RFE = 3	64.49%
Model 2: RFE = 5	73.18%
Model 3: RFE = 7	74.63%
Model 4: RFE = 8	76.08%
Model 5: RFE = 9	76.81%
Model 6: RFE = 10	77.53%
Model 7: RFE = 11	77.53%
Model 8: RFE = 12	78.26%

Table 2.1.2: Comparison

2.1.2 Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a method of feature selection that works by fitting a model and then choosing the best and worst features. A user can specify the number of features by setting the RFE value. In Table 2.1.2 you can see a side-by-side view of 8 models where the number of features is increased to the full 12 (after target and dropping one feature in preprocessing). In this case, all 12 are needed as you can see the Random Forest model accuracy increase. However, this is the processed data and so the features are more important, different results might happen when the unprocessed 76 features are used.

An RFE model with 3 features tells us that the most important features are; Chest Pain Type, Serum Cholesterol and Max Heart Rate. These three also had the highest positive and negative values in the correlation matrix which shows both methods are showing the same features that have the most influence.

An attempt was made to try this same method on the full unprocessed data with 76 features but did not work as there is a much larger amount of missing data and it would take a considerable amount of time to process this data. However, the idea behind this method is sound and would be the basis for next steps on improving the performance of the ML models.

2.1.3 Feature Importance with Random Forest

Random Forest (RF) models are some of the most popular and perform very well in several tasks. RF models have several built-in techniques that can be used for feature Selection. When a RF model has completed it can rank the performance of the features in the model. Figure 2.1.3 shows a plot of which features were more important. The top 3 are the same as we saw in the previous two methods. However, this time with values assigned to the features we can see that age is closely related to the top 3. While this method is related to RF models its use of values allows us to gain a greater insight into how each feature is important.

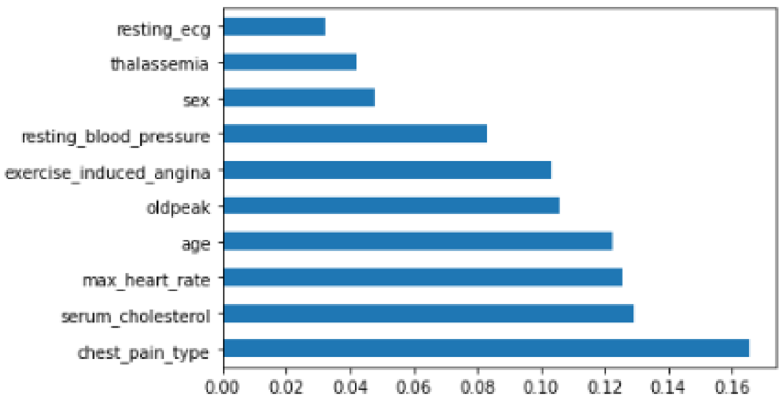

Figure 2.1.3: Plot of Random Forest features

Feature	Score
Chest Pain Type	262.72
Exercise Induced Angina	212.56
Max Heart Rate	129.39
Oldpeak	98.29
Sex	95.71
Thalassemia	81.86
Age	79.73
Slope	40.44
Serum Cholesterol	39.71

Table 2.1.4: Values

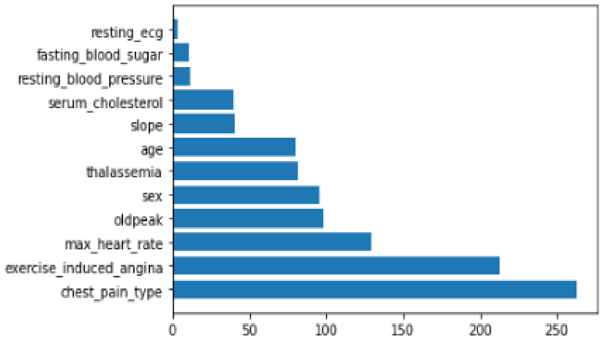

Figure 2.1.4: Plot of values

2.1.4 Univariate Selection

Univariate selection is a statistical method that can be used to see what features have the strongest relationship by examining the strength between the feature and the target variable. Different methods can be used to give a value, such as chi-squared, ANOVA, etc. It's a simple method to run and intercept with a downside that it does not consider the potential effects between features. Table 2.1.4 shows the features using "f_classif" as the scoring metric. As we can see in Figure 2.1.4, again we have the same top 3 features with the highest scores. What is interesting is after this Oldpeak and sex come a close fourth and age is much further down the list.

3. Methodology

3.1. Preprocessing Steps

The first step in preprocessing was to combine the four datasets into one. Next we had to deal with missing values.

Missing values

Combining the datasets can lead to several missing values. In total there are 920 observations with Table 3.1 showing the amount of missing values in each feature.

Feature Name	Missing Values
Age	0
Sex	0
Chest pain type	0
Resting blood pressure	59
Serum cholesterol	30
Fasting blood sugar	90
Resting ECG	2
Max heart rate	55
Exercise-induced angina	55
Oldpeak	62
Slope	309
Major	611
Thalassemia	486
Target	0

Table 3.1: Feature names and missing value count

There is no one type of fix to deal with all types of missing data. Different techniques need to be used to deal with missing data as different features require different approaches.

- **Average value by another column:** Resting Blood Pressure, Serum Cholesterol, Max Heart Rate, and Oldpeak are values that change with age. So, taking an overall average would skew the results, the research took an average value of each by age category. This approach leads to a better performance with the models.
- **Replace value with the Mode value:** Fasting Blood Sugar, Resting ECG, Exercise-induced Angina and Slope all have a low amount of category values, so the mode value was taken to replace the missing values.
- **Replace with a Normal value:** Thalassemia has a normal range of 3, since there was no way to know what the missing values are, the normal value was used to replace the missing values.
- **Drop column option:** Major had 611 values missing out of 920. This is a very high amount so the research decided to drop this column. Future work has been noted to research this in more detail to increase domain knowledge to understand the best approach that could be used instead of dropping it.

3.2. Motivation for the models

For the research, Logistic Regression, Random Forest and Gradient Boost Models were used. The Logistic Regression model was used as a baseline as it is widely used in binary classification problems and it offers a simple and interpretable model that estimates the outcome based on a linear combination of the features [18]. Random Forest is an ensemble learning method that combines multiple Decision Trees to make predictions and has been showed to perform well for heart disease data [19]. It is very useful when dealing with high dimensional data, so it was selected to first be used on the processed data and then on the larger unprocessed data if this was possible. Lastly, Gradient Boosting was selected as another ensemble technique as it builds a sequence of weak prediction models in a wide range and it has been known to create very good performance results for prediction models [20]. Gradient boosting models can also handle both categorical and contusions features and was also chosen to see how they would compare with the processed and unprocessed data.

Using the three models the research was looking at taking advantage of each of their strengths. Logistic Regression gives interpretability, Random Forest can capture complex relationships and Gradient Boosting can take many weaker predictions to get a strong prediction.

3.3. Model Tunning with Hyperparameters

For each of the models, a Grid Search CV method was used to try and improve the model's performance by using a range of tuning values.

Logistic Regression Grid Search Parameters:

C: Regularization parameters tested with values of 0.001, 0.01, 0.1, 1, 10, 100. Penalty: "l1" and "l2". Solver: Two types were used with "lbfgs" and "saga". "lbfgs" was paired with "l2" and "saga" was used with both "l1" and "l2".

Random Forest Grid Search Parameters:

N_estimators: Values tested with a range from 100 to 2000 in steps of 100. Max_features: Four strategies were used, "auto", "sqrt", "log2" and None. Max_depth: Values tested range from 10 to 100 in steps of 10, as well as None. Min_sample_split: tested with 2,5,10. Min_sample_split: values of 1,2,4 tested. Bootstrap: True and False Parameters.

Gradient Boosting Grid Search Parameters:

N_estimators: Values tested with a range from 100 to 2000 in steps of 100. Learning_rate: 0.001, 0.01, 0.1 and 0.2. Max_depth: values of 3, 4, 5, and 6 were used. Min_sample_split: tested with 2,5,10. Min_sample_split: values of 1,2,4 tested. Subsample: 0.5, 0.8 and 1 were tested.

4. Evaluation

4.1. Model Performance

As this is a medical research project and looks to predict heart disease, in medical diagnostics, it's typically more important to look at correctly identifying the disease case (class 1) even at the expense of the false positives (class 0). Because the cost of a false negative of missing a case of heart disease is often much greater than a false positive, which would mean additional tests but no negative health outcome for the patient. The below metrics in Table 4.1 are for class 1 in this case.

Models	**Test Data**			
	Accuracy	**Precision**	**Recall**	**F1-Score**
Logistic Regression	78%	80%	79%	80%
LG: Grid Search - Best model	78%	80%	79%	80%
Randon Forest – Best Model Overall	**80%**	**83%**	**81%**	**82%**
RF: Grid Search - Best Model	80%	83%	81%	82%
Gradient Boosting	78%	81%	78%	79%
GB: Grid Search - Best Model	77%	80%	78%	79%

Table 4.1: Comparison of results on Class 1 for test data

4.2. Compare Model Performances

From Table 4.1, all the models performed with similar results in terms of the metrics on Class 1. With Logistic Regression the default model had an accuracy of 78% and the best model with Grid Search had the same performance metrics which suggested that the Grid Search did not find any combination of hyperparameters that would improve over the default settings. Overall Random Forest performed the best with an accuracy of 80% and once again using Grid Search did not improve the overall performance. For Gradient Boosting the best model performed better than the Grid Search tuning with an accuracy of 78%.

Overall, the models performed relatively well with the best model being the default Random Forest. It was surprising to see that the Grid Search Models either performed the same or worse than the default models (this does depend on what the parameters are) and this is an area that was been marked for future research. Other parameters in the tuning can we used to explore different combinations of parameters.

5. Conclusions and Future Work

5.1. Conclusion

This research has successfully demonstrated the potential for Machine Learning models to be used with healthcare data to help in the prediction of heart disease. This research used Logistic Regression, Random Forest and Gradient Boosting models to take the combined data of 4 separate datasets and achieve an overall accuracy of 80% on the test dataset with Radom Forest performing the best.

This research also started to explore the aim of improving on this performance by using the full unprocessed data of 76 attributes and investigated 4 Feature Selection methods; Heatmap with Correlation, Recursive Feature Elimination, Feature Importance with Random Forest and Univariate Selection that could be used on this data to improve on the model's performance. Using these methods on the processed data highlighted Chest Pain Type, Exercise Induced Angina and Max Heart Rate as the most significant features when predicting heart disease. These features were consistent across all 4 Feature Selection techniques.

This research also stressed the importance of practical applications of the best-performing models with the creation of a simple user interface which then used the best-performing model to give a user a prediction of heart disease. This takes the project beyond the research and shows how it can be used in a real-world application for the early detection of heart disease and how it could also be used as an example of using these types of prediction models to help with the reduction of waiting times.

5.2. Challenges of Giving Medical Advice with Machine Learning

Artificial intelligence research is a growing area but so far there has been limited application across the various domains of medicine [21]. As seen in our research the best result we had was an 80% accuracy on the test data and in the medical world, this is not high enough. The safe and timely transition of AI models from research needs to be thoroughly researched, clinically evaluated and go above and beyond in order to offer a quality of care for patients' outcomes. A research paper that offers the 12 key challenges [22] in medical machine learning and sums up the challenges include; ensuring data is of sufficient quality, establishing baseline performance standards, and performance across models and populations.

5.3. Future Work

Using Grid Search did not give better performance results for Logistic Regression or Random Forest and even a small decrease for Gradient Boosting. This was unexpected and I would like to explore this in more detail with other values for the hyperparameters as it would be generally expected for Grid Search to improve a model's performance.

With any medical prediction, accuracy is very important and a score of 80% found in this research needs to be improved. The next step is to continue to work on the 76 attributes and apply preprocessing steps on these and then use the Feature Selection methods to select the best features and look at improving the accuracy score for the lowest number of features.

In hindsight, using two ensemble models was too similar and next I would have liked to have tried other Machine Learning Models. Previous research has also shown very good results from SVM's and Neural Networks, so another step is to use these to look at improving the model's performance, first with the processed 14 features and then with the 76 attributes once they have been improved.

Lastly, this research created a small application to show the model in use to give a user advice. The next step would be to add in more error control and then create a website or app that would allow users to use this more easily. An overall aim would be to create a suite of tests that can be easily used by users to help highlight any areas where an improvement in lifestyle could lead to a lower risk of health conditions.

6. References

[1] "RBC Capital Markets | The healthcare data explosion." https://www.rbccm.com/en/gib/healthcare/episode/the_healthcare_data_explosion (accessed May 16, 2023).

[2] V. Gain, "Irish software Smartlist aims to reduce hospital waiting lists by 25pc," *Silicon Republic*, Sep. 27, 2021. https://www.siliconrepublic.com/enterprise/yellowschedule-vitro-smartlist-software-hospital-waiting-lists (accessed May 16, 2023).

[3] E. Handley, "Using AI, analytics and machine learning to improve NHS wait times," *Open Access Government*, Feb. 13, 2023. https://www.openaccessgovernment.org/using-ai-analytics-machine-learning-improve-nhs-wait-times/153084/ (accessed May 16, 2023).

[4] "ireland modern health care system - Google Search." https://www.google.com/search?q=ireland+modern+health+care+system&rlz=1C1CHBF_en-GBIE1043IE1044&oq=ireland+modern+health+care+system&aqs=chrome..69i57j33i10i160j33i22i29i30.5767j0j7&sourceid=chrome&ie=UTF-8 (accessed May 16, 2023).

[5] "Janet Keane: Cancer misdiagnosis lawsuits and cases," *Irish Legal News*, Nov. 29, 2022. https://www.irishlegal.com/articles/janet-keane-cancer-misdiagnosis-lawsuits-and-cases (accessed May 18, 2023).

[6] "Access, Waiting Times and Expenditure on Healthcare," *Social Justice Ireland*, Apr. 25, 2022. https://www.socialjustice.ie/article/access-waiting-times-and-expenditure-healthcare (accessed May 16, 2023).

[7] "Private health expenditure in Ireland: Assessing the affordability of private financing of health care - ScienceDirect." https://www.sciencedirect.com/science/article/abs/pii/S0168851019301861?via%3Dihub (accessed May 16, 2023).

[8] "Cardiovascular diseases." https://www.who.int/health-topics/cardiovascular-diseases (accessed May 17, 2023).

[9] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, "Computational intelligence for heart disease diagnosis: A medical knowledge driven approach," *Expert Syst. Appl.*, vol. 40, no. 1, pp. 96–104, Jan. 2013, doi: 10.1016/j.eswa.2012.07.032.

[10] "UCI Machine Learning Repository: Heart Disease Data Set." https://archive.ics.uci.edu/ml/datasets/heart+disease (accessed May 18, 2023).

[11] "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization | SpringerLink." https://link.springer.com/article/10.1023/A:1007607513941 (accessed May 18, 2023).

[12] Y. Freund and L. Mason, "The Alternating Decision Tree Learning Algorithm," in *Proceedings of the Sixteenth International Conference on Machine Learning*, in ICML '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Jun. 1999, pp. 124–133.

[13] E. N. Smirnov, I. G. Sprinkhuizen-Kuyper, and G. I. Nalbantov, "Unanimous Voting using Support Vector Machines," *Unanimous Voting Using Support Vector Mach.*, 2005.

[14] A. Almulihi *et al.*, "Ensemble Learning Based on Hybrid Deep Learning Model for Heart Disease Early Prediction," *Diagnostics*, vol. 12, no. 12, p. 3215, Dec. 2022, doi: 10.3390/diagnostics12123215.

[15] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert Syst Appl*, vol. 36, pp. 7675–7680, May 2009, doi: 10.1016/j.eswa.2008.09.013.

[16] R. R. Sarra, A. M. Dinar, M. A. Mohammed, and K. H. Abdulkareem, "Enhanced Heart Disease Prediction Based on Machine Learning and χ2 Statistical Optimal Feature Selection Model," *Designs*, vol. 6, no. 5, Art. no. 5, Oct. 2022, doi: 10.3390/designs6050087.

[17] S. Arooj, S. ur Rehman, A. Imran, A. Almuhaimeed, A. K. Alzahrani, and A. Alzahrani, "A Deep Convolutional Neural Network for the Early Detection of Heart Disease," *Biomedicines*, vol. 10, no. 11, p. 2796, Nov. 2022, doi: 10.3390/biomedicines10112796.

[18] A. G, B. Ganesh, A. Ganesh, C. Srinivas, Dhanraj, and K. Mensinkal, "Logistic regression technique for prediction of cardiovascular disease," *Glob. Transit. Proc.*, vol. 3, no. 1, pp. 127–130, Jun. 2022, doi: 10.1016/j.gltp.2022.04.008.

[19] M. Pal and S. Parija, "Prediction of Heart Diseases using Random Forest," *J. Phys. Conf. Ser.*, vol. 1817, no. 1, p. 012009, Mar. 2021, doi: 10.1088/1742-6596/1817/1/012009.

[20] K. L. Kumar and B. E. Reddy, "Heart Disease Detection System Using Gradient Boosting Technique," in *2021 International Conference on Computing Sciences (ICCS)*, Dec. 2021, pp. 228–233. doi: 10.1109/ICCS54944.2021.00052.

[21] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med.*, vol. 17, no. 1, p. 195, Oct. 2019, doi: 10.1186/s12916-019-1426-2.

[22] R. J. Ellis, R. M. Sander, and A. Limon, "Twelve key challenges in medical machine learning and solutions," *Intell.-Based Med.*, vol. 6, p. 100068, Jan. 2022, doi: 10.1016/j.ibmed.2022.100068.