International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

A Comparative Overview of Machine Learning Models for Heart Disease Diagnosis

Ashna Jain (TISB)[1], Dhruv Roongta (TISB)[2]

Abstract: ***Heart Disease is the leading cause of death in the United States, with one out of every 4 deaths being linked to it. A recent study revealed that 40% of the participating middle-age adults had coronary heart disease without being aware of it, making early-stage diagnosis extremely important. [1]To combat human error, Machine Learning has become increasingly involved in the process of medical diagnosis, with a standard ML Algorithm outperforming 72% of doctors[2]. Moreover, researchers have found that 5% of patients in the USA are annually misdiagnosed, which can be potentially fatal. Several Algorithms have been tested to find the optimal accuracy of heart disease diagnosis, with Support-Vector Machines (SVM) and Decision Trees being the most accurate. This paper further introduces SVM Ensembles and Random Forests as a result of Ensemble Learning often being more accurate than single models due to reduced bias and overfitting, as found by (Opitz and Maclin, 1999). We evaluate each of the four models to find the most accurate model for medical diagnosis, particularly detection of heart disease. We then use this to calculate the best way to lower a patient's probability of developing heart disease in the next five years, analysing how simple lifestyle changes can make a significant difference. The results portray that the SVM Ensemble is the most accurate, depicting its potential in medical diagnosis.***

Keywords: heart disease prediction, machine learning models, SVM ensemble, medical diagnosis, random forest

1. Introduction

Medical diagnosis is a key application and often used to test the limits of Machine Learning models. A central reason behind this is the importance of improving the accuracy of medical diagnosis, specifically in regard to reducing the number of False Negatives. While doctors have been fairly accurate with identifying diseases, individual biases, lapses of judgements, or other factors can impact their accuracy.

Resultantly, Machine Learning is increasingly used as a 'co-pilot', and helps for more accurate diagnosis. (Richens et al., 2009) finds that a standard Machine Learning model is more accurate than 72% of doctors. This represents an important shift in the field of healthcare, clearly portraying that Artificial Intelligence is the future, emphasising the need for research into factors that determine model accuracy.

The majority of Machine Learning Research focuses on analysing how models react to various changes, such as missing data or noise, which circumstances are suited to certain models, testing out novel methods, et cetera. Our paper employs a similar approach. With regard to heart disease detection, prior research has primarily focused on Support Vector Machines, Decision Trees, and Random Forests. (Hassani et al., 2020) finds that, for heart disease prediction, the SVM had the highest accuracy with 94.60%, followed by the Random Forest and Decision Tree with 89.20% and 89.10% respectively.

Our paper analyses the SVM, SVM Ensemble, Decision Tree, and Random Forest as a result of the high accuracies that the SVM, Decision Tree and Random Forest present in the context of medical diagnosis, and the promising findings that papers on the SVM Ensemble, such as (Stork et al., 2015; Claesan et al., 2014; Mohareb et al., 2016) present.

The primary purpose of this paper is to analyse the 4 previously mentioned models in relation to heart disease prediction, specifically focusing on the SVM Ensemble and whether its construction consisting of individual Support Vector Machines acts to a considerable advantage, and to further understand the specific features that make certain models better at classification. We further test whether ensemble learning is better for medical diagnosis due to the reduction of individual weaknesses.

The following is an overview of the paper: We first analyse the SVM, SVM Ensemble, Decision Tree and Random Forest individually in Section 2. We also analyse the data set that we have chosen, in Section 3. This information is then used to construct a hypothesis in Section 4, which is tested out in Section 5, studying the results and how parameter tuning affected the models. Section 6 presents the 5-year prediction model we construct. Finally, a conclusion is contained in Section 7.

2. Section 2

This section provides an overview of the four machine learning models that we have selected, the SVM, SVM Ensemble, Decision Tree, and Random Forest, studying how they are built, their advantages and disadvantages, and how well they may be suited to medical diagnosis, specifically heart disease.

2.1 SVM

The Support-Vector Machine, designed by Vapnik and his co-workers at AT&T Laboratories (Vapnik, 1998) (Cortes and Vapnik, 1995) has grown to become one of the most accurate Machine Learning models, with applications in medical diagnosis (Veropoulos et al., 1999), time series prediction ((Fernandez, 1999) and more. Built as a linear classification, supervised learning model, the SVM creates a binary

[1] *Early heart disease detection saves lives*. (2021). Providence. https://www.providence.org/news/uf/644387305
[2] Richens et al., (August 2020),*Improving the accuracy of medical diagnosis with causal machine learning, https://www.nature.com/articles/s41467-020-17419-7#Sec3*

separator such that it has the maximum possible margin. For nonlinear classification tasks the Support Vector Machine uses the kernel trick to transform the data into a higher dimension such that a linear hyperplane can separate it.

Advantages:
(Tian et al., 2011) summarises the success of the SVM to the principle of maximal margin, dual theory and the kernel trick, three key components of the Support Vector Machine. The paper also mentions how the SVM reduces most classification problems to optimization problems, at which it excels.

The principle of maximal margin is a key differentiator between the SVM and other Machine Learning models, and an integral influence on its accuracy. One way the maximal margin hyperplane helps the SVM is through its relation to the intuitive method humans would use to linearly separate two classes. The maximal-margin also reduces the likelihood of the SVM overfitting (Vapnik 1998).

The Dual Problem of the SVM allows it to be more efficient than through the form of a primal problem as it can be solved through quadratic programming algorithms. This is a significant factor in the SVM's speed and energy-efficiency. The simplified model can be represented as:

$$\text{maximize } f(c_1 \dots c_n) = \sum_{i=1}^{n} c_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i c_i (\mathbf{x}_i^T \mathbf{x}_j) y_j c_j,$$

$$\text{subject to } \sum_{i=1}^{n} c_i y_i = 0, \text{ and } 0 \leq c_i \leq \frac{1}{2n\lambda} \text{ for all } i.$$

[3]

Lastly, the kernel trick is likely the leading factor in the SVM's accuracy, allowing it to transform data into a higher hyperplane such that it is still linearly classifiable. Taking the dot-product also allows the SVM to be computationally cheaper.

Disadvantages:
Despite being one of the most accurate models, the SVM has several disadvantages. To begin with, the SVM is prone to overfit when the number of features is greater than the number of samples. While this is not the case for our data set, other fields of medical diagnoses might be subject to it.

(Han and Jiang, 2014) finds that the SVM is more likely to encounter overfitting when using a non-linear kernel, as is the case with the majority of data. This may also hamper the accuracy of the SVM. Combined with the SVM's 'weak spot' with noise, in certain cases the SVM may offer a high train accuracy but a significantly lower test accuracy. However, we suspect that this is not the case with our data set.

Selecting an appropriate kernel function can also be a challenging task. This poses a challenge in finding the highest accuracy and constructing the ideal model for varying situations.

2.2 SVM Ensemble

The SVM Ensemble is a machine learning model composed of individual SVMs that are each trained on a subset of the original training sample. These aggregated SVM models, similar to the decision trees in the Random Forest, provide the ensemble the strength of the individual models collectively while surpassing individual bias.

Advantages
Ensemble methods often have various advantages over the singular model, as discussed by (Kuncheva and Whitaker, 2003; Sollich and Krogh, 1995).

To begin with, an advantage of using the SVM Ensemble is the decrease in training time. (Claesen et al., 2015) finds that, counterintuitively, despite the increased number of models that have to be trained, the SVM ensemble reduces the computational effort required. While SVMs are largely computationally cheap, large data sets can make them expensive in certain instances.

On a similar note, the SVM Ensemble, like most ensemble methods, is better suited than the SVM for large data sets. This may however not largely influence our circumstance as a result of the relatively small data set.

More importantly, the SVM Ensemble is found to be as or more accurate than singular SVMs. (Stork et al., 2015) finds that the Ensemble reduces bias and overfitting. While the SVM is relatively well-suited to be prone to overfitting, as discussed in the previous section, if the number of variables is larger than the number of samples overfitting may take place. The Ensemble variant effectively counters this, while having further pros in the case of non-linear kernels.

Using several kernels can help increase accuracy and decrease overfitting, as individual bias is reduced, while the specialty of different kernels can be used together to achieve a model that is better suited to a wider range of data samples.

Disadvantages
While Ensemble methods can improve upon individual models, they may also pose disadvantages in certain cases. This most notably refers to when individual SVMs are not sufficiently weak enough and resultantly the ensemble

[3] Figure obtained from https://en.wikipedia.org/wiki/Support-vector_machine

overfits. Finding the ideal training sample size to train the individual nodes on can be a challenging process as trial-and-error may need to be employed.

However, on the whole, the SVM Ensemble has a strong set of advantages that allow it to improve on individual errors that SVMs may make, and through the use of different kernels and boosting can likely increase the accuracy. A notable portrayal of this is (Huang et al.,2017) which shows that the SVM Ensemble outperforms the SVM in breast cancer prediction, a medical diagnosis similar to heart disease prediction.

2.3 Decision Tree

The Decision Tree is a popular machine learning algorithm that was first invented in 1963 by the department of statistics at the University of Wisconsin-Madison. It represents the data as a tree, with each internal node of the representation denoting an attribute and each leaf node a class label. It can be used to solve both classification and regression problems.

Advantages
Advantages of the decision tree algorithm include that data processing is easier, as data normalisation or scaling is not required. This reduces the time needed to prepare the data and train the model, making it suitable for real-time analysis, as well as rendering it computationally cheap.

Missing data points and outliers also do not considerably impact the process, and decision trees perform very well on large datasets, so they are not prone to large biases.

Disadvantages
Relatively small changes in data can cause much larger changes in the decision tree's structure, rendering it unstable. The decision tree's calculation complexity, time required for training, and cost is typically higher than that of other machine learning models. They can also overfit, and this becomes more of a problem when a tree is deep. One way to combat this is to set a maximum depth but this leads to error due to bias. Hence it is difficult to create a tree which doesn't overfit and has a high level of accuracy.

Decision Trees work by breaking data sets into smaller and smaller subsets, while incrementally developing the associated decision tree. However due to their tendency to overfit, they are not the best option for medical diagnosis. Past research done also found this algorithm only showed superior accuracy 33 % of the time when compared with other machine learning models in diagnosing diseases including diabetes, Parkinson's disease and breast cancer following that it likely wouldn't be the most accurate for heart disease either.

2.4 Random Forest

The Random Forest model is a supervised learning technique. It is a decision tree ensemble which takes a number of decision trees and takes the average to improve the predictive accuracy of a dataset. An ensemble is a machine model that combines predictions from two or more models, which can make better predictions and achieve better performance than any single contributing model. The greater number of trees leads to greater accuracy and also solves the problem of overfitting that individual decision trees face. This ability to limit overfitting while not increasing error due to bias is why it is such a powerful model.

Advantages
The advantages include that the Random Forest takes less training time and has a higher accuracy when compared to other models. Moreover, it is able to maintain a high accuracy even when a large proportion of data is missing, which is very important in medical diagnosis. The Random Forest algorithm outperformed every other machine learning model in a study which compared the accuracy of different supervised learning algorithms in disease detection, showing superior accuracy 53 % of the times.

It is also capable of performing both regression and classification tasks and can handle datasets with high dimensionality.

This ensemble algorithm has many benefits over a single decision tree. They outperform decision trees as features are chosen randomly during the training process, allowing them to generalize over the data in a better way. This gives them a higher accuracy and with the heart disease dataset we obtained an accuracy of 83 percent. The Random Forest is much more robust than a single decision tree, and yields more useful results.

However, while they can be used for both regression and classification tasks, they are not as suitable for regression tasks. This is because it can't extrapolate and can only make a prediction that is the average of previously observed labels. Since we need continuous data as output, regression must be used.

Disadvantages
While the Random Forest is a powerful and accurate model, it has certain disadvantages that make it unsuitable to some data sets. A primary disadvantage is its larger computational complexity that increases the training time, making it less effective for real-time prediction. This is however not a large problem with heart disease prediction.

Another problem is the apparent 'black box' nature of the model, not allowing notable insight into how parameters are evaluated, making parameter tuning more complicated. This could be a key disadvantage in medical diagnosis as learning how features are weighted could be valuable.

On the whole the Random Forest is an accurate model that is likely a top contender for medical diagnosis.

3. Dataset Analysis

This section evaluates the dataset that we have chosen, the Cleveland dataset[4], which contains 14 features and 270 rows. While this is relatively small, collecting data for medical diagnosis is difficult due to privacy concerns and legal challenges. However (Althian et. al., 2011) finds that "dataset size is not necessarily an obstacle to a high performing model". This indicates that the size of the dataset is an unlikely factor in the accuracies of the models.

Some of the parameters in the dataset include age, blood pressure, and cholesterol, which have been identified as the leading causes for heart disease alongside lifestyle factors like smoking habits, and hence this dataset can accurately be used for diagnosis and prediction. (*Know Your Risk for Heart Disease Cdc.Gov*, 2019)

While 14 features is a large number of features, we notice that they are largely uncorrelated, and their inclusion hence presents the models with a 'richer' dataset making them less likely to overfit.

We find that our dataset does not have any missing data points, which is advantageous as replacing missing data points with averages can distort models and create biases.

4. Hypothesis

Through our analysis of the SVM, SVM Ensemble, Decision Tree, and Random Forest in Section 2 and our identification of the properties of the dataset in Section 3, we formulate a hypothesis regarding which model is the most accurate for heart disease prediction in this section.

Our selection of machine learning models features two distinct types: singular and ensemble methods. Ensemble methods construct a new model by aggregating a set of individual models, with (Opitz and Maclin, 1999) finding that they often provide a higher accuracy due to the removal of individual bias and overfitting. (Polikar, 2006) provides evidence of Ensemble methods often, intuitively, reducing training time, making analysis of large data sets more efficient. They are also proven useful when the data set is sufficiently small by training individual models on a subset.

Resultantly, we narrow it down to the SVM Ensemble and the Random Forest, both of which are highly accurate Machine Learning models. We use regression as a result of working with continuous data. The limitation of Random Forests in regression is that they cannot extrapolate and can only predict values within the training dataset[5]. Moreover, the SVM Ensemble is also better suited for heart disease (and medical diagnosis in general) as a result of its effectiveness in binary classification, whereas the Random Forest is more effective at multi-class classification. The collective result of these parameters is that we hypothesise the SVM Ensemble is likely to perform better in medical diagnosis.

The SVM Ensemble itself has two varieties to transform the collection of individual SVMs into an effective Ensemble Algorithm: bagging and boosting. The primary difference between bagging and boosting is the sequential training that boosting employs as a result of certain data samples being weighted. (Stork et al., 2015) finds that bagging is better than for larger sample sizes, while boosting performs better on smaller training samples. (Huang et al., 2017) finds similar results, discovering that boosting generally outperforms bagging. In our case, the data set is considerably small enough for boosting to be preferable. Hence, we predict that the SVM ensemble with boosting will be the most accurate machine learning model for prediction of heart disease and medical diagnosis in general.

5. Results

The results support our hypothesis, demonstrating that the SVM Ensemble had the highest accuracy on our data set[6], followed by the Random Forest, SVM and Decision Tree. (See Figure 1)

We created the ensemble using both bagging and boosting, and a significantly higher accuracy using boosting, reinforcing the view that boosting is preferable for smaller datasets like ours. We also observed increased randomness with bagging, which in the case of predicting heart disease and medical diagnosis as a whole, is not desirable. (See Figure 2)

In terms of future work, a major area that we may explore is how different kernel numbers affect the accuracy of the SVM model and hence the ensemble. Moreover, the number of False Negatives is higher than the number of False Positives in our model, which in the future we would reduce. We would also test the SVM Ensemble on other datasets to help further look into its strengths and weaknesses.

Moreover, asides from further optimising the SVM Ensemble to obtain even higher accuracies, we also want to apply it to diagnosing other diseases such as diabetes and cancer. Analysing other machine learning models could further provide insight into which features and parameters yield high accuracies in medical diagnosis. Eventually we also want to apply the SVM Ensemble to other fields such as stock market prediction or spam mail identification.

6. Probability

[4] https://data.world/informatics-edu/heart-disease-prediction/workspace/file?filename=+Heart_Disease_Prediction.csv

[5] Thompson, B. (2019, December 20). *A limitation of Random Forest Regression - Towards Data Science*. Medium. https://towardsdatascience.com/a-limitation-of-random-forest-regression-db8ed7419e9f#:%7E:text=It%20can%20only%20make%20a,labels%20in%20the%20training%20data.

Mwiti, D. (2021, May 25). *Random Forest Regression: When Does It Fail and Why?* Neptune.Ai. https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why

[6] https://data.world/informatics-edu/heart-disease-prediction/workspace/file?filename=+Heart_Disease_Prediction.csv

In this section, we create a machine learning model to predict the probability of somebody developing heart disease over the course of the next five years. Since we found the SVM ensemble to be the most accurate machine learning model for medical diagnosis in section 5, we will base our model on that, using it to calculate the initial probability of developing heart disease. This is important as cardiovascular disease remains the main cause of mortality and hence early diagnosis which can lead to timely treatment is of utmost importance and save lives. (Capotosto et al., 2018)

Asides from alerting a patient to a higher possibility of them developing heart disease and urging them to get medical attention, this model will also evaluate the main factor that is leading to this heightened probability along with links to sources on how to make lifestyle changes to lower this probability. This will be based on the factor entered that is furthest from the optimum value. It will also return the reduced probability of developing heart disease in the next 5 years if that change is made, encouraging the patient to see what a significant difference lifestyle changes can make. This is in no way a substitute to professional medical advice but rather a calculation of probability. The main goal is to alert people of the urgency of a doctor's appointment and not to replace it entirely.

The initial probability is calculated using the SVM ensemble and then 5 is added to the age factor to calculate the changed probability in 5 years if no other factor changes. Then the main factor found to be furthest from the optimum value is calculated. As of now the main factors considered are cholesterol and BP, however in the future we would expand this range as we get bigger datasets.

The factor that is furthest from the optimum is identified as the primary factor relating to the risk of heart disease. The model constructs a hypothetical where this factor is at the optimum to calculate the impact that this factor contributes in terms of probability.

7. Conclusion

Our paper compares the SVM, SVM Ensemble, Decision Tree, and Random Forest, hypothesising that the SVM Ensemble is best suited for heart disease diagnosis. We find that the results match the hypothesis, with the SVM Ensemble having the highest accuracy of 84.44%. We evaluate the use of a machine learning ensemble and find that it will usually outperform a single model in medical diagnosis. Boosting is also shown to be more accurate than bagging, and the SVM to be more reliable and accurate than a Decision Tree. The SVM ensemble, while still relatively novel, has portrayed its advantages over the SVM in several tests, and we believe this paper acts as a strong proof of the potential that it has in medical diagnosis.

References

[1] Evgeniou, T., Pontil, M. (2001), Support Vector Machines: Theory and Ap- plications, https://www.researchgate.net/publication/221621494 Support Vector Machines - Theory and Applications
[2] Tian, Y.; Shi, Y.; Liu, X. 2012. Recent advances on support vector machines research, Technological and Economic Development of Economy 18(1): 5–33
[3] Karthiga, Mary, Yogasini, A. S. K. S. M. M. Y. (2017). Early Prediction of Heart Disease Using Decision Tree Algorithm. Research Gate. Published. https://www.researchgate.net/publication/315023624 Early Prediction of Heart - Disease Using Decision Tree Algorithm
[4] Uddin, S., Khan, A., Hossain, M. E., Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19(1). https://doi.org/10.1186/s12911-019-1004-8
[5] K, D. (2020, December 26). Top 5 advantages and disadvantages of Decision Tree Algorithm. Medium. https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision- tree-algorithm-428ebd199d9a
[6] Sharma, A. (2020, May 12). Decision Tree vs. Random Forest – Which Algo- rithm Should you Use? Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest- algorithm/
[7] Machine Learning Random Forest Algorithm - Javatpoint. (2011). Www.Javatpoint.Com. https://www.javatpoint.com/machine-learning-random-forest-algorithm
[8] Brownlee, J. (2021, April 26). Why Use Ensemble Learning? Machine Learn- ing Mastery. https://machinelearningmastery.com/why-use-ensemble-learning/
[9] Heart Disease — cdc.gov. (2021b, January 19). Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/index.html
[10] Early heart disease detection saves lives. (2021b). Providence. https://www.providence.org/news/uf/644387305
[11] Thomas, J. (2020b, July 16). Heart Disease: Facts, Statistics, and You. Healthline. https://www.healthline.com/health/heart-disease/statisticsWho-is-at-risk?
[12] Leibowitz, D. (2020b, October 15). AI Now Diagnoses Disease Better Than Your Doctor, Study Finds. Medium. https://towardsdatascience.com/ai-diagnoses-disease-better-than-your-doctor-study- finds-a5cc0ffbf32
[13] SVM Ensembles Are Better When Different Kernel Types Are Combined. (2015). Research Gate. Published.://www.researchgate.net/publication/300896635 SVM Ensembles Are Better When - Different Kernel Types Are Combined
[14] Huang, M. W., Chen, C. W., Lin, W. C., Ke, S. W., Tsai, C. F. (2017). SVM and SVM Ensembles in Breast Cancer Prediction. PLOS ONE, 12(1), e0161501. https://doi.org/10.1371/journal.pone.0161501
[15] Patil, M. (2019). Prediction and Analysis of Heart Disease Using SVM Algo- rithm. International Journal for Research in Applied Science and Engineering Technology, 7(1), 856–859. https://doi.org/10.22214/ijraset.2019.1137
[16] Sandhy, Y. A. (2020). Prediction of Heart Diseases using Support Vector Machine. International Journal for

Research in Applied Science and Engineer- ing Technology, 8(2), 126–135. https://doi.org/10.22214/ijraset.2020.2021

[17] Mohareb, F., Papadopoulou, O., Panagou, E., Nychas, G. J., Bessant, C. (2016). Ensemble-based support vector machine classifiers as an efficient tool for quality assessment of beef fillets from electronic nose data. Analytical Methods, 8(18), 3711–3721. https://doi.org/10.1039/c6ay00147e