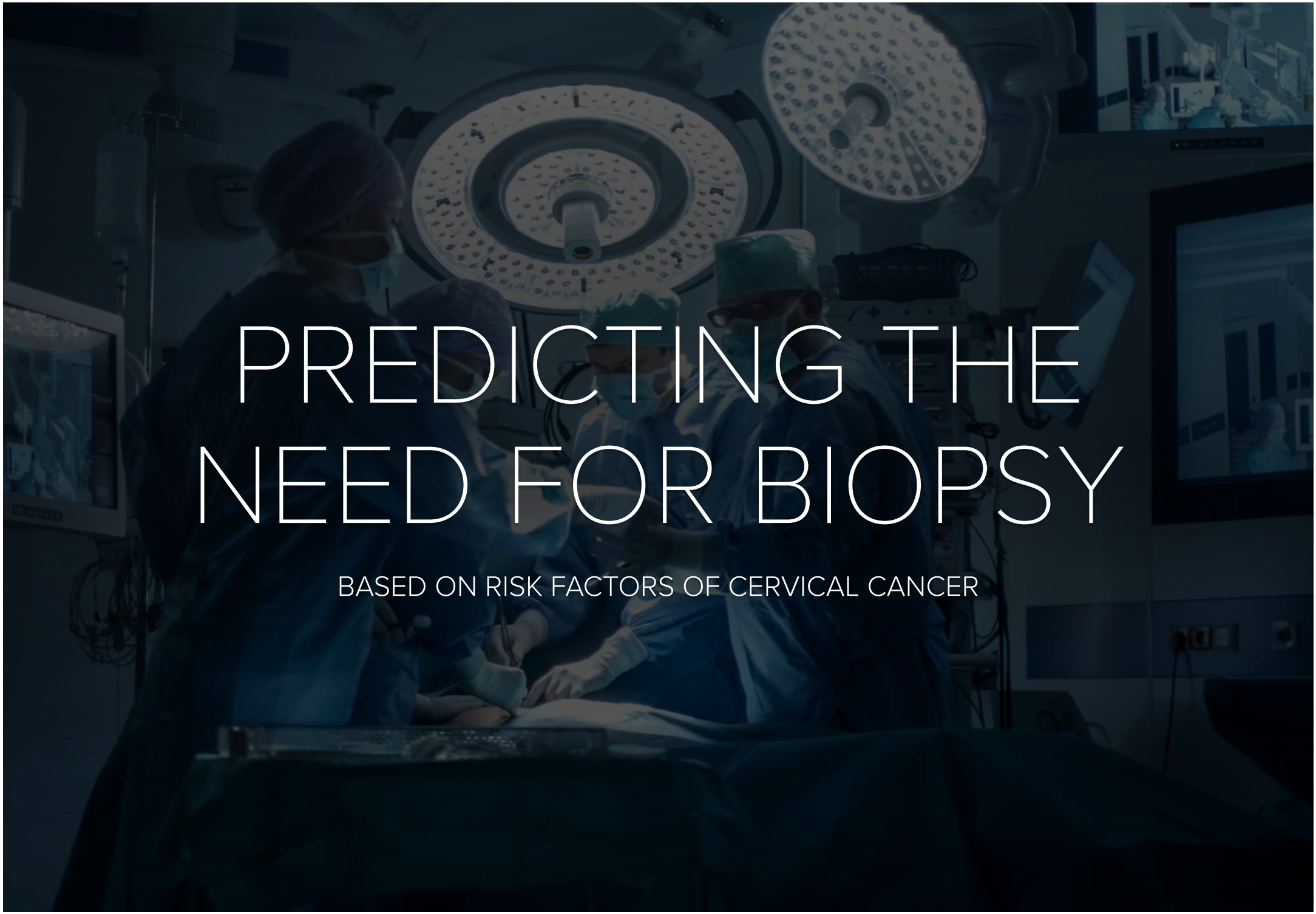


PREDICTING THE NEED FOR BIOPSY

BASED ON RISK FACTORS OF CERVICAL CANCER

About **11,000 new cases** of invasive cervical cancer are diagnosed each year in the U.S. However, the number of new cervical cancer cases has been declining steadily over the past decades. Although it is the most preventable type of cancer, each year **cervical cancer kills about 4,000 women in the U.S.** and about **300,000 women worldwide**. In the United States, cervical cancer mortality rates plunged by 74% from 1955 - 1992 thanks to increased screening and early detection with the Pap test, a method of cervical screening used to detect potentially precancerous and cancerous processes in the cervix or colon. However, we believe that with the technological advancements of machine and deep learning, we can make it easier for medical services to detect cervical cancer more efficiently so patients can undergo biopsy in the earlier stage.

AIM: TO BUILD EFFECTIVE PREDICTIVE MODELS THAT CAN ACCURATELY CLASSIFY WHETHER OR NOT THE PATIENT NEEDS TO UNDERGO CERVICAL BIOPSY

ABOUT DATASET

This dataset is obtained from UCI Repository. The dataset contains patients from Hospital Universitario de Caracas in Caracas, Venezuela that exhibit risk factors that lead to cervical cancer. Moreover, the dataset originally contains 36 columns (features) which comprise demographic information, habits, and historic medical records of 858 patients. It is also noted that several patients decided not to answer some of the questions because of privacy concerns. The target of this dataset is identified. The target's column name is "Biopsy", meaning the final outcome is whether or not this patient should undergo a biopsy.

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	...	STDs: Time since first diagnosis
0	18	4.0	15.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	?
1	15	1.0	14.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	?
2	34	1.0	?	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	?
3	52	5.0	16.0	4.0	1.0	37.0	37.0	1.0	3.0	0.0	...	?
4	46	3.0	21.0	4.0	0.0	0.0	0.0	1.0	15.0	0.0	...	?
...	...	...	...	...	...	...	...	...	...	...	...	...
853	34	3.0	18.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	?
854	32	2.0	19.0	1.0	0.0	0.0	0.0	1.0	8.0	0.0	...	?
855	25	2.0	17.0	0.0	0.0	0.0	0.0	1.0	0.08	0.0	...	?
856	33	2.0	24.0	2.0	0.0	0.0	0.0	1.0	0.08	0.0	...	?
857	29	2.0	20.0	1.0	0.0	0.0	0.0	1.0	0.5	0.0	...	?

Table 1: Overview of the dataset (not all columns are displayed)

Download Dataset

DATA CLEANING

1. It can be seen that this dataset uses '?' to represent null values. Therefore, we will **replace '?' with 'NaN'** to ease the processing.
2. After replacing with null, we **count** how many null values there are in each column of the dataset.
3. From the above step, it can be observed that the parameters "**STD: Time since the first diagnosis**" and "**STD: Time since last diagnoses**" had many null values. Replacing these null values would make the classifier useless. Hence, these two features were dropped for each training, validation and test set.
4. Next, for the other features that contain null values, we will use **imputation**.
5. Since there are multiple columns that have the 'object' data type, the values in the columns for each set were **converted to numerical values** to ease the processing.
6. The columns in the dataset have both categorical and numerical attributes. Hence, we **separate the categorical and numerical attributes** for each set as they both require different processes/formulas in imputing the null values.

7. **Descriptive statistics** was used to replace the missing values. The most typical metrics for this task are mean, median and mode. The **median was used to replaced numerical attributes**, and **categorical attributes were replaced by the mode**. Mean value imputation was avoided since they highly influence the extreme values/outliers in the data.

DATA PREPROCESSING

Step 1: Feature Scaling/Normalization

Feature scaling is a method used to normalize the range of independent variables or features of data. In this process, a **standard scalar** was used to normalize the data. This is because it uses the standard normal distribution. All the means of the attributes are made zero, and the variance is scaled to one. As our dataset has both numerical and categorical variables, we will only normalize the numerical attributes.

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	...	STDs:Hepatitis B	STDs:HPV
46	1.080830	-0.311264	0.360061	0.530241	0	-0.295999	-0.202021	1	0.831613	0	...	0	0
41	1.198575	-0.311264	0.360061	-0.183961	0	-0.295999	-0.202021	0	-0.570927	0	...	0	0
405	-1.156307	-0.311264	-0.356719	-0.898163	0	-0.295999	-0.202021	1	-0.430673	0	...	0	0
199	0.021134	-0.311264	0.360061	-0.183961	0	-0.295999	-0.202021	0	-0.570927	0	...	0	0
340	-0.920819	0.297083	0.360061	-0.898163	0	-0.295999	-0.202021	1	-0.290419	0	...	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
158	0.609854	-0.311264	1.435231	-0.183961	0	-0.295999	-0.202021	0	-0.570927	0	...	0	0
132	0.374366	-0.311264	1.793621	-0.183961	0	-0.295999	-0.202021	1	-0.548486	0	...	0	0
621	0.492110	-0.919610	-0.715109	1.958644	0	-0.295999	-0.202021	1	-0.150165	0	...	0	0
150	0.138878	-0.311264	0.360061	-0.183961	0	-0.295999	-0.202021	1	0.270597	0	...	0	0
234	-0.214355	-0.311264	0.360061	-0.183961	0	-0.295999	-0.202021	1	-0.430673	0	...	0	0

Table 2: Dataset with normalised values using standard scalar

Step 2: Feature selection using Pearson's correlation technique

Pearson's correlation feature selection technique was utilized to find redundant features. This feature selection technique compares the degree of association among all variables. When there is a high correlation between two independent attributes, one of these attributes can be removed since both features contribute the same to the ML model. By observing the diagonal values, any variable that is directly correlated to itself will show a positive correlation. Therefore, age has a positive correlation, which is one, and so the diagonal should also be visible. The dark color shows a near-zero correlation. This technique can only be used on numerical attributes.

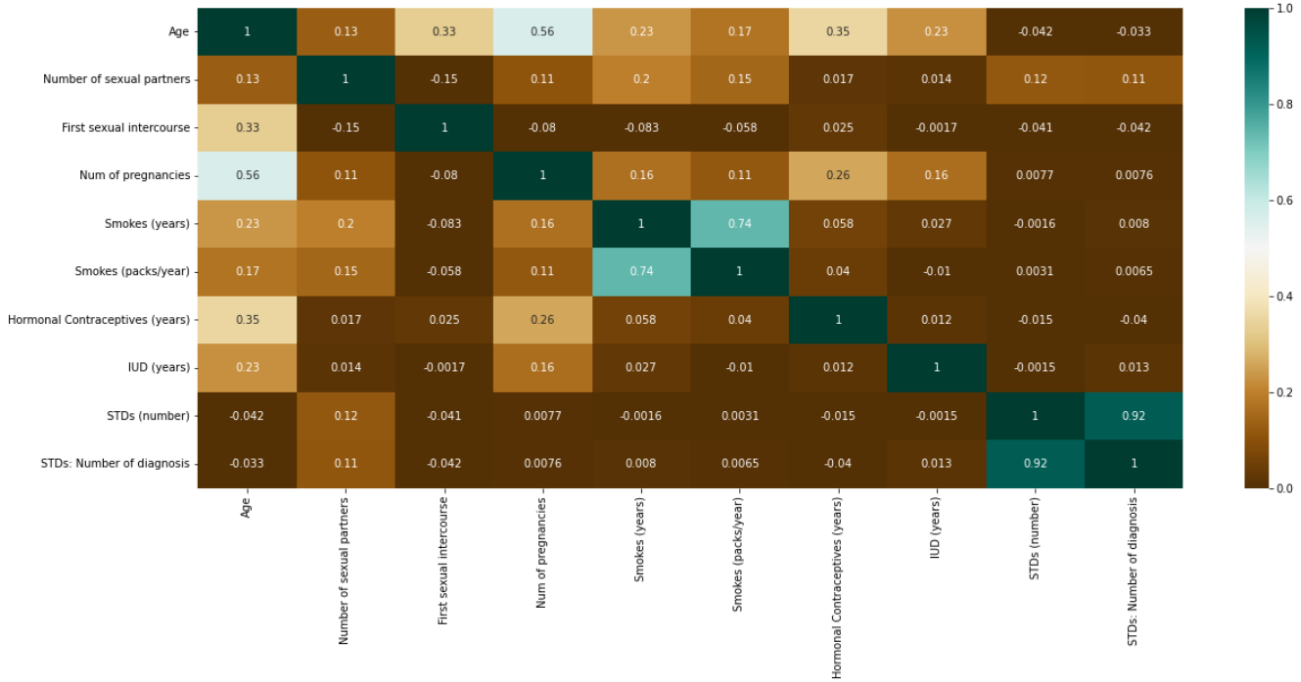


Figure 1: Pearson's correlation heatmap

If the correlation value is above 0.8, we consider that is highly correlated hence the feature is removed. From the heatmap, we can observe that the feature "**STDs (number)**" has a value of 0.92 so this feature was removed.

Step 3: Feature selection using the feature importance with Tree Based Classifier

The importance of each feature is determined by using a Tree-Based Classifier, namely the Extra Trees Classifier. The normalized total reduction

in the mathematical criteria used in the decision of the feature of the split is computed. This value is called the Gini Importance of the feature. Based on the previous step, we have already selected the best features for numerical attributes. For this step, we will use this technique to select the best features for categorical attributes.

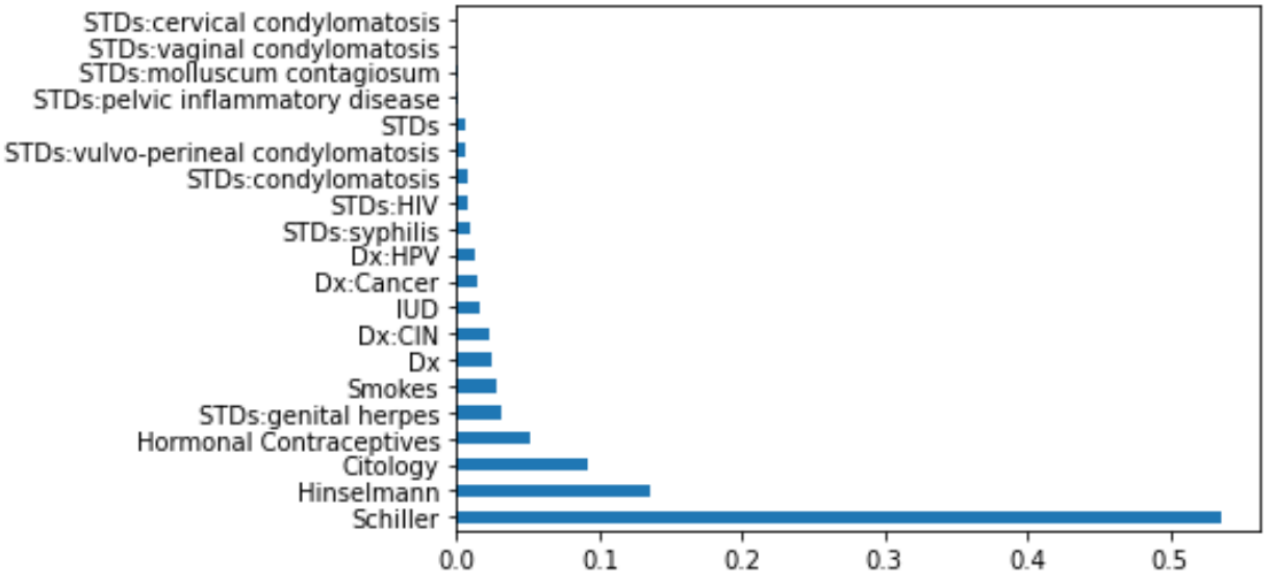


Figure 2: Bar graph of the feature importances

From the graph above, we can observe that there are four features that show no importance hence these features are dropped.

The result of the finalised selected features also aligns with the scientific point of view. According to the American Cancer Society, several risk

factors that lead to cervical cancer are HPV, HIV, sexual history, smoking, and long-term use of hormonal contraceptives.

DATA MODELLING

Three predictive models are built using the **Neural Network** and **Hybrid Intelligent System** algorithms. The ratio of the split is 55% training set, 25% validation set, and 20% test set.

Model 1: Neural Network

Neural networks, also known as artificial neural networks or simply neural nets, are a type of machine learning model inspired by the structure and functioning of the human brain. They consist of interconnected artificial neurons organized into layers. Each neuron receives input signals, performs a computation, and passes the output to the next layer until the final output is produced.

1. Building the neural network

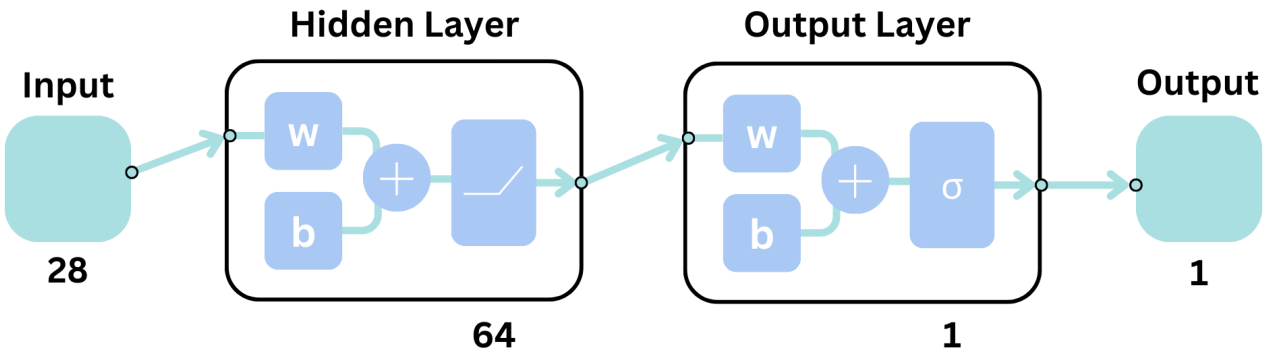


Figure 3: Overview of the Neural Network model

- Our neural network model has 28 input features, which serve as the initial information that is fed into the network. This model is a sequential neural network. The input features are then connected to a hidden layer, which contains 64 neurons. The choice of 64 neurons in the hidden layer is based on experimentation and finding the optimal balance between model complexity and performance to fine-tune this parameter.
- In our model, the activation function used in the hidden layer is the rectified linear unit (ReLU). ReLU is a popular activation function that introduces non-linearity to the model. It helps the network learn complex relationships and can prevent the problem of vanishing gradients during training.
- The output layer of our neural network employs the sigmoid activation function. Since our dataset is binary classification, sigmoid activation is

commonly used to map the output to a probability between 0 and 1. It allows the network to predict the probability of the input belonging to the positive class.

2. Regularisation techniques

- To prevent overfitting and improve generalization, our model incorporates L2 regularization. L2 regularization, also known as weight decay, adds a penalty term to the loss function based on the squared magnitude of the weights. This encourages the network to find a simpler and smoother solution by discouraging large weights.
- Dropout regularization is applied with a rate of 0.1. Dropout randomly sets a fraction of input units to zero during training, which helps prevent overreliance on specific features or neurons. It encourages the network to learn more robust and generalized representations.
- Our model also utilizes early stopping regularization with a parameter called patience set to 3. Early stopping is a technique that monitors the model's performance on a validation set during training. If the performance does not improve for a certain number of epochs, training is stopped to prevent overfitting and find the best-performing model.

3. Performance Metrics and Confusion Matrix

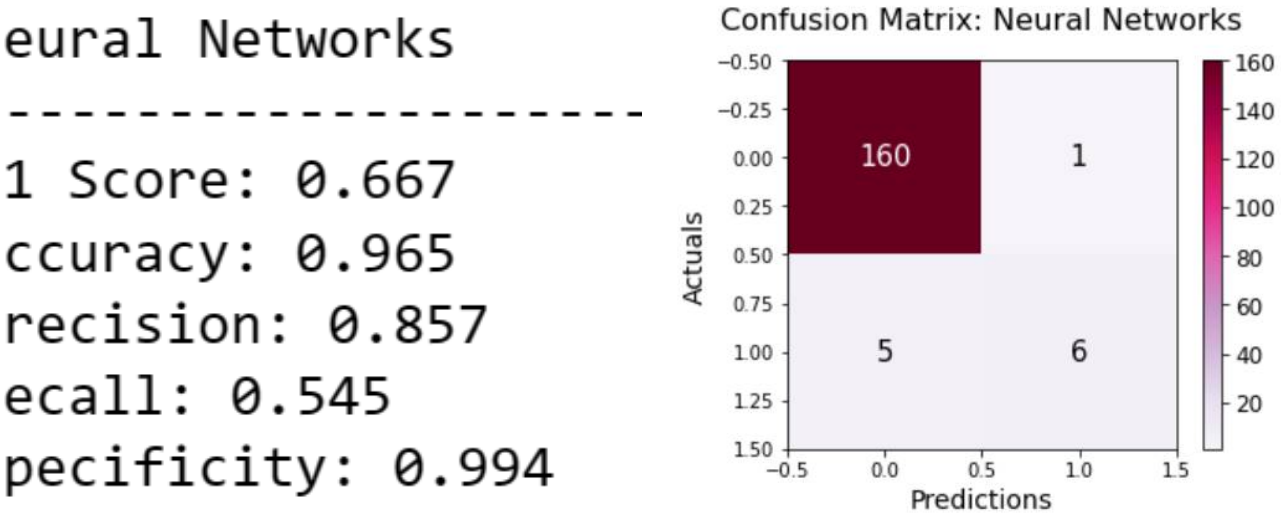


Figure 4: Performance metrics and confusion matrix using the Neural Network model

Based on both the figures above, we can observe that

- The **F1 score** is a measure of the model's accuracy, combining both precision and recall. It is the harmonic mean of precision and recall. In our case, the F1 score is **0.667**, indicating a reasonable balance between precision and recall. A higher F1 score indicates better performance, with 1 being the best possible score.

- **Accuracy** represents the proportion of correctly classified instances out of the total instances. In our case, the accuracy is **0.965**, which indicates that the model is able to classify the majority of instances correctly.
- **Precision** measures the proportion of true positive predictions out of the total predicted positives. In our case, the precision is **0.857**, indicating that when the model predicts a positive class, it is correct **85.7%** of the time. High precision suggests a low rate of false positives.
- **Recall**, also known as sensitivity or true positive rate, measures the proportion of true positive predictions out of the actual positive instances. In our case, the recall is **0.545**, indicating that the model correctly identifies **54.5%** of the positive instances. High recall suggests a low rate of false negatives.
- **Specificity**, also known as true negative rate, measures the proportion of true negative predictions out of the actual negative instances. In our case, the specificity is **0.994**, indicating that the model correctly identifies **99.4%** of the negative instances. High specificity suggests a low rate of false positives.
- Overall, the neural network model shows good accuracy and high specificity, indicating a strong ability to correctly classify negative instances. However, there is room for improvement in terms of

precision, recall, and F1 score, especially in correctly identifying positive instances.

Model 2: Hybrid Intelligent (Decision Tree + Genetic Algorithm)

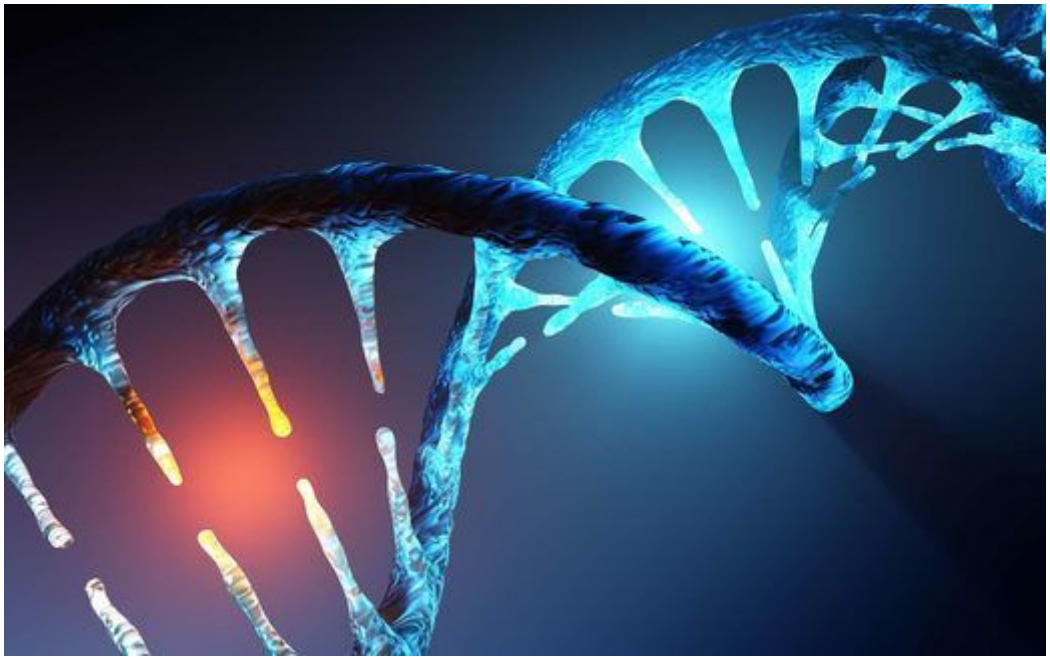


Figure 6: Genetic Algorithm

Genetic Algorithms (GA) are evolutionary algorithms that mimic natural selection to generate and evolve a population of candidate solutions. They use genetic operators like selection, crossover, and mutation to explore different solutions.

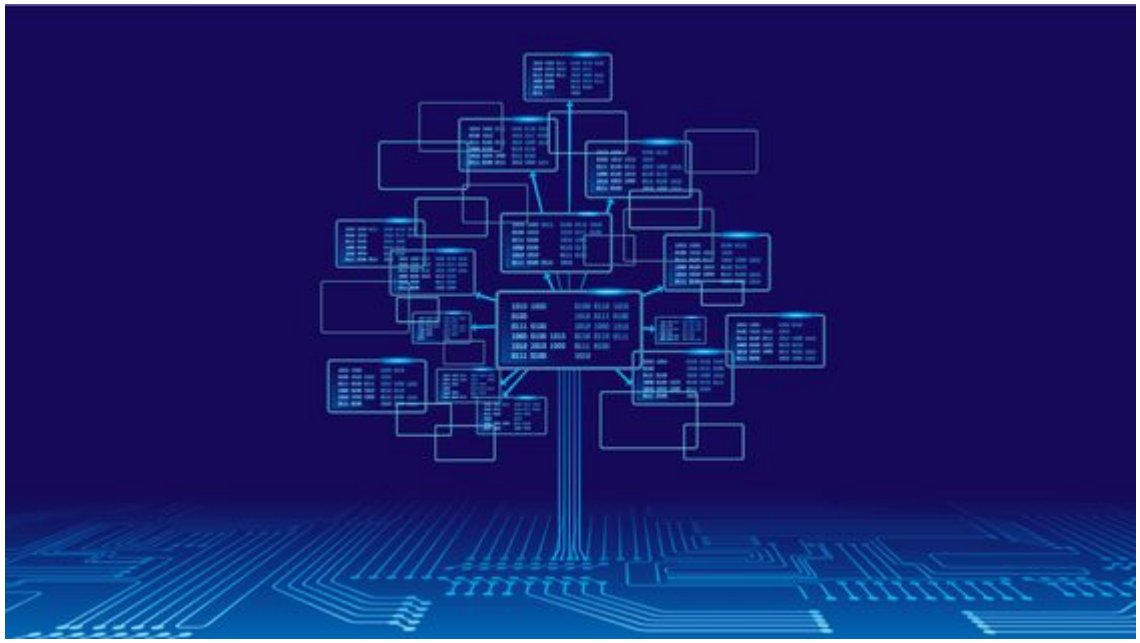


Figure 7: Decision Tree

On the other hand, **Decision Trees (DT)** are machine learning models that create a tree-like structure to make decisions based on features in a dataset. They are commonly used for classification tasks.

By **combining** the **global search** capabilities of **GA** with the **local optimization** capabilities of **DT**, the hybrid approach **enhances the accuracy** and **performance** of the DT model in predicting the Biopsy target. This allows for more accurate and reliable predictions, aiding in the analysis and understanding of the dataset related to Biopsy prediction.

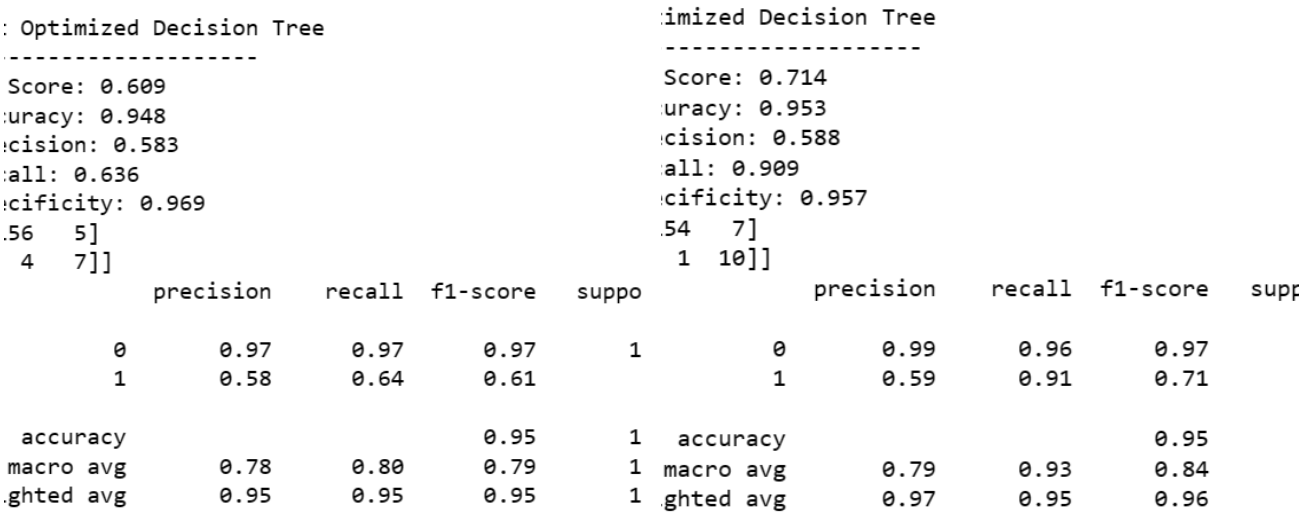
In Genetic Algorithm (GA), the population size, number of generations, crossover probability, and mutation probability are important parameters that influence the behavior and performance of the algorithm and the optimization of our DT model. Hence, for our hybrid approach, we have set the parameters to significant levels:

- **Population Size**: The population size determines the number of individuals (candidate solutions) in each generation. We set the population size at a moderately sized population of **80** as we do not only aim to increase the demographic variety of the population and develop the best solutions but also use less computational cost of evaluation and evolving of the population.
- **Number of Generations**: The number of generations specifies how many iterations the GA will go through. Each generation represents a complete cycle of evaluating, selecting, recombining, and mutating the population. In our optimisation, we set the number of generations at **20** to allow more iterations for the algorithm to converge towards better solutions but still minimizing the computation time.
- **Crossover Probability**: Crossover is the genetic operator that combines genetic material from two parent individuals to create offspring. The crossover probability determines the likelihood of

performing crossover for a pair of parents. In our model, we set the crossover probability at **0.8** to maintain the diversity and accelerate the convergence to good solutions.

- **Mutation Probability**: Mutation is the genetic operator that introduces random changes into an individual's genetic material. It allows the exploration of new regions in the solution space that may not be accessible through crossover alone. By setting the mutation probability at **0.2**, it indicates a relatively low probability. A lower mutation probability ensures that the **algorithm focuses more on recombining existing genetic material through crossover,** while still allowing some level of exploration.

The optimized DT model obtained through the hybrid approach is expected to provide improved accuracy compared to a non-optimized DT model.



: Optimized Decision Tree

Score: 0.609
uracy: 0.948
cision: 0.583
all: 0.636
cificity: 0.969
56 5]
4 7]]

	precision	recall	f1-score	suppo
0	0.97	0.97	0.97	1
1	0.58	0.64	0.61	
accuracy			0.95	1
macro avg	0.78	0.80	0.79	1
ghted avg	0.95	0.95	0.95	1

imized Decision Tree

Score: 0.714
uracy: 0.953
cision: 0.588
all: 0.909
cificity: 0.957
54 7]
1 10]]

	precision	recall	f1-score	supp
0	0.99	0.96	0.97	
1	0.59	0.91	0.71	
accuracy			0.95	
macro avg	0.79	0.93	0.84	
ghted avg	0.97	0.95	0.96	

Figure 8: Comparisons of performance metrics before and after optimization

Based on the results that we obtained from Figure 8, we can conclude that:

- **F1 score**: The non-optimized Decision Tree model achieves an F1 score of **0.609**, indicating a moderate balance. However, the optimized model shows a **significant improvement**, with an F1 score of **0.714**, demonstrating a better balance between correctly identifying positive cases and minimizing false negatives and false positives.
- **Accuracy**: The non-optimized model achieves an accuracy of **0.948**, indicating that it correctly predicts 94.8% of the instances. In

comparison, the optimized model achieves a higher accuracy of **0.953**, demonstrating an i**mproved overall prediction accuracy**.

- **Precision**: The non-optimized model has a precision of **0.583**, suggesting that only 58.3% of the predicted positive cases are true positives. In contrast, the optimized model shows an improved precision of **0.588**, indicating that approximately 58.8% of the predicted positive cases are true positives. This reduction in false positive predictions demonstrates an **enhancement in the precision** of the optimized model.
- **Recall**: The non-optimized model achieves a recall of **0.636**, capturing 63.6% of the actual positive cases. In contrast, the optimized model achieves a **significantly improved** recall of **0.909**, accurately identifying 90.9% of the actual positive cases. This improvement reflects the optimized model's enhanced ability to identify positive cases.
- **Specificity**: The non-optimized model demonstrates a specificity of **0.969**, accurately identifying negative cases with a high accuracy of 96.9%. The optimized model **maintains a similarly high** specificity of **0.957**, indicating its ability to correctly identify negative cases with a high accuracy of 95.7%. This suggests that the optimization process did not compromise the model's ability to correctly identify negative cases.

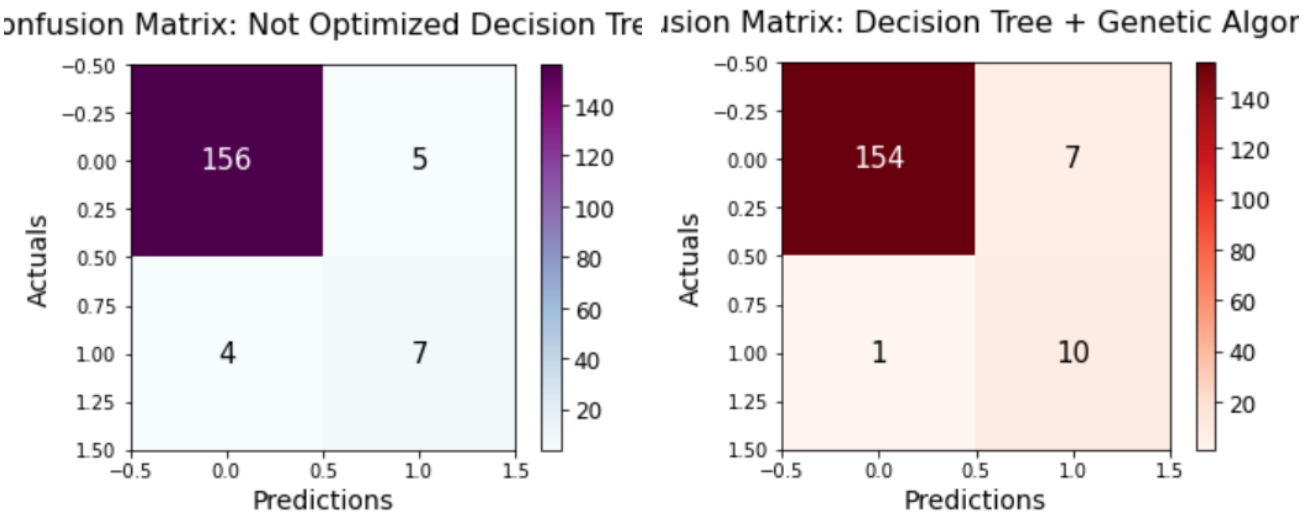


Figure 9: Comparisons of confusion matrix before and after optimization

From the Confusion Matrix shown above, we can conclude that: In the non-optimized Decision Tree model, the confusion matrix reveals **156** true negatives (TN), **5** false positives (FP), **4** false negatives (FN), and **7** true positives (TP). This means that the model correctly predicted 156 instances as negative and 7 instances as positive, while making 5 false positive predictions and 4 false negative predictions.

However, in the optimized Decision Tree model, there is a noticeable improvement with **154** true negatives (TN), **7** false positives (FP), **1** false negative (FN), and **10** true positives (TP). This indicates that the optimized model has achieved a better ability to correctly classify both negative and positive cases. It correctly predicted 154 instances as negative and 10 instances as positive, while making 7 false positive predictions and only 1 false negative prediction.

The **reduction in false negatives** and **false positives** in the optimized model signifies an **improvement** in its performance. It demonstrates the model's enhanced accuracy in correctly classifying instances and its ability to minimize both types of errors. This improvement aligns with the higher accuracy, precision, recall, and F1 score observed in the optimized Decision Tree model. Overall, the **optimized Decision Tree model shows better predictive capabilities** and a more **reliable** classification performance, as reflected in the improved confusion matrix results.

In conclusion, the optimized Decision Tree model exhibits notable improvements across various evaluation metrics compared to the non-optimized model. It achieves a better balance between precision and recall, higher accuracy, increased precision, improved recall, and maintains

a high specificity. These enhancements highlight the **effectiveness of the hybrid approach** in **optimizing** the Decision Tree model for predicting the Biopsy outcome.

MODEL EVALUATION

As displayed and evaluated via the explanation using both models above, the justifications are as below:

- The **F1 score** using the Neural Network model **(0.667)** is **lower** than of the Optimized Decision Tree model **(0.714)**. Since the closer the value of F1 score to the value 1, denotes a better classifier, hence the Optimized Decision Tree model performs better in this case. In this case, based on the formulae, F1-score is more sensitive to false negative and will penalize models that produce too many false negatives as compared to accuracy.
- The **accuracy** of both models are **slightly similar** with the Neural Network model having a **slightly higher** accuracy of **0.956** while the Optimized Decision Tree has an accuracy of **0.953** which shows that they both have a relatively high value of accuracy.

- The **precision** using the Neural Network model **(0.857)** is **higher** than that of the Optimized Decision Tree **(0.588)**. Even though the higher precision values denote a better performance of algorithm, but in this case, this metric is not significantly important, as when the model incorrectly labels as positive that are actually negative, hence the person who will not be needing biopsy will be at risk of receiving one.
- The **recall** using Neural Network model **(0.545)** is lower than that of the Optimized Decision Tree **(0.909)**. Recall is the measure of the model correctly identifying True Positives whereby it predicts correctly the patients who needs a biopsy as compared to the actual scenario. Hence, the Optimized Decision Tree works best in predicting the patients and their actualness of needing a biopsy.
- The specificity using Neural Network model **(0.994)** is **slightly higher** than that of the Optimized Decision Tree **(0.957)**. Specificity mentions about how negative records are correctly predicted. Hence, with the high specificity of both the models, it will help in evaluating which patient does not need a biopsy correctly.

Among all the performance metrics, **F1-score** should be prioritized to compare and determine the most optimal algorithm in this dataset. To justify this decision, firstly, accuracy is not suitable to be used because it is

only optimal for classes that are balanced and there is no serious flaw in predicting false negatives (FN). In this case, the **Optimized Decision Tree** model has a higher value of F1-score of **0.714** compared to the Neural Network model with a value of **0.667**. Additionally, the Decision Tree model also has the lowest value of False Negatives. This is significant when related to predicting a sickness of an individual. As an example, if a patient is suffering from cervical cancer, but the trained model predicts that the patient does need to undergo a biopsy, it would result in a false negative which is bad as the sick patient is predicted to be healthy and ends up not receiving cervical biopsy. Therefore, the more false negatives, ultimately it will cause more patients to not undergo a biopsy when needed.

In conclusion, the Optimized Decision Tree model using Hybrid Intelligence is our champion model.