International Journal of Healthcare Information Systems and Informatics
Volume 16 • Issue 4

Post Thoracic Surgery Life Expectancy Prediction Using Machine Learning

Akshaya Ravichandran, EY GDS, India
Krutika Mahulikar, Uniphore, India
Shreya Agarwal, Maynooth University, Ireland
Suresh Sankaranarayanan, SRM Institute of Science and Technology, Chennai, India
https://orcid.org/0000-0001-5145-510X

ABSTRACT

Lung cancer survival rate is very limited post-surgery irrespective of if it is small cell or non-small cell. A lot of work has been carried out by employing machine learning in life expectancy prediction post thoracic surgery for patients with lung cancer. Many machine learning models like multi-layer perceptron (MLP), SVM, naïve Bayes, decision tree, random forest, logistic regression have been applied for post thoracic surgery life expectancy prediction based on data sets from UCI. Also, work has been carried out towards attribute ranking and selection in performing better in improving prediction accuracy with machine learning algorithms. So accordingly, the authors, here, have developed a deep neural network-based approach in prediction of post thoracic life expectancy which is the most advanced form of neural networks. This is based on dataset obtained from Wroclaw Thoracic Surgery Centre machine learning repository which contained 470 instances. On comparing the accuracy, the results indicate that the deep neural network can be efficiently used for predicting the life expectancy.

KEYWORDS

KNN, MLP, Naïve Bayes, NSCLC, SVM, UCI

1. INTRODUCTION

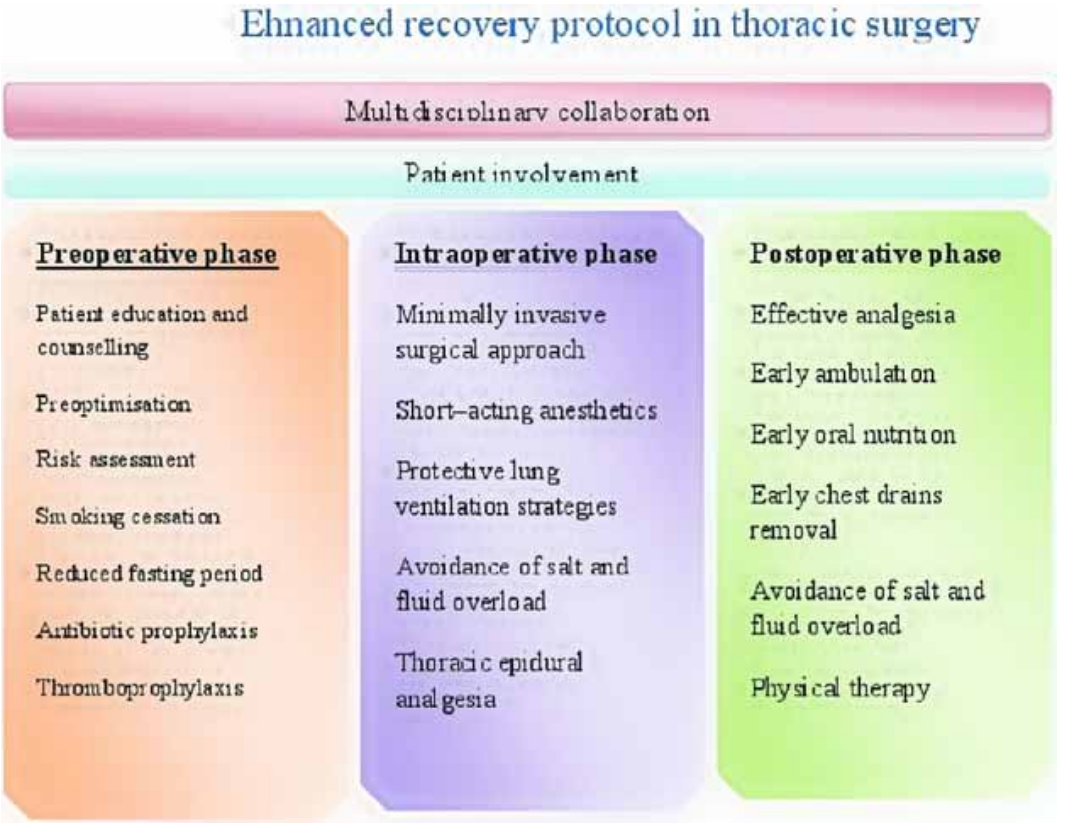
Touted as the leading cause of "cancer" death across the globe, lung cancer has been among the most common type of malignancies diagnosed on adults (Ferlay, et al., 2010; Sigel, et al., 2020). Aiding decisions in operative, perioperative, and/or surgical thoracic procedures, researchers such as Desuky & El Bakrawy (2016) and Danjuma (2015) have evaluated the performance of machine learning (ML) algorithms, for example, multilayer perceptron (MLP), J48, and the Naive Bayes (NB), on the University of California Irvine (UCI) Machine Learning (ML) repository thoracic surgery dataset.

As **Figure 1** shows, thoracic surgery may be split into three (3) specialties: (1) adult cardiac surgery; (2) congenital or pediatric heart surgery; and (3) general thoracic surgery.

In operative, perioperative, and/or surgical critical care of patients (American Medical Association, n.d.) who obtained congenital pathologic conditions within the chest, thoracic surgery is often recommended. Even so, recent studies have predicted that around 80% of lung cancer patients are diagnosed with non-small cell lung cancer (NSCLC) and around 25% with early-stage operable disease (e.g., Adam, et al., 2014; Timmerman, et al., 2016; Sarna, et al, 2008). For the NSCLC early stages, the preferred treatment has been curative lung resection. Symptoms for NSCLC include pain,

DOI: 10.4018/IJHISI.20211001.oa32

Figure 1. Consequence of Thoracic Surgery about here



fatigue, the decay of lung function, cardiorespiratory fitness and quality of life. Similar to NSCLC, small-cell lung cancer (SCLC) is very aggressive.

Notwithstanding, with the low post-surgery survival rate of lung cancer patients whether it is SCLC or NSCLC (e.g., Timmerman et al, 2016; Adam et al, 2019), many critical factors such as age, experience of the surgeon and patient medical condition, among other things, must be considered in determining the risk of operating on these patients. Hence, a thorough diagnosis and analysis must be performed based on past historic patient data and current patient medical condition prior to recommending surgery. With respect to the prediction of post-thoracic life expectancy, there has been emerging work implementing ML techniques. Predictions from the use of these techniques are often good enough to assist the patient (and surgeon) in deciding to undergo surgery or not.

Given the limited cancer patient survival rate post-thoracic surgery, research has emerged in applying data mining techniques for its medical diagnosis and prediction (e.g., Nachev & Reapy, 2015). Models used included decision trees, Naïve Bayes (NB), artificial neural network (ANN) and support vector machines (SVM). More recently, Desuky & El Bakrawy (2016) have applied MLP, Naïve Bayes, J48, logistic regression (LR) for post-thoracic surgery life expectancy prediction on the UCI thoracic surgery dataset. Their work also involved attribute ranking and selection to achieve more accurate prediction. While only satisfactory accuracy has been achieved with traditional ML algorithms as well as with more recent ones such as Multi-Layer Perceptron (MLP), a type of Artificial Neural Network (ANN), and Bayesian model, no work on deep learning (DL) for post-thoracic life expectancy has yet been found.

Deep Neural Network (DNN) which is a part of Deep Learning (DL) is an advancement of ANN; if applied, it stands to achieve better accuracy with a reduced percentage of error vis-à-vis other ML algorithms (Geron, 2016). As it is touted to be superior in performance vis-à-vis other traditional ML algorithms such as LR, SVM, ANN, Decision Tree (DT), and Random Forest (RF), we hereby proposed to develop a DNN-based approach, the most advanced form of NNs in ML, to predict post-thoracic life expectancy. In this work, the thoracic surgery dataset was drawn from the Wroclaw Thoracic Surgery Centre ML repository, which contained 470 instances.

The rest of the paper is organized as follows. Section 2 reviews background information from the extant literature for ML in post-thoracic surgery prediction on life expectancy. Section 3 shifts focus to our proposed DL work with specific insights on the theoretical contributions of different ML algorithms with insights on use of data flow and use case diagrams whereas Section 4 reports on the results and analysis of the proposed work. Finally, Section 5 concludes with summary remarks and highlights potential future work.

2. BACKGROUND

Thoracic surgery is considered the consummating operation being performed on carcinoma patients. Survival rate (kokulu, et al., 2015) is a key factor for surgeons to determine on which patient surgery would be beneficially performed. Patient selection is one of the challenging factors in thoracic surgery decision, taking into account parameters to determine risk-benefit considerations for the patient both in the short-term (e.g. post-operative complications, including death-rate within the ðrst month) and long-term perspective (e.g. survival for 1-5 years).

In the last decades, different ML algorithms have been studied as well as evaluating attribute ranking and selection methods towards disease prognosis and prediction. Zieba, et al. (2014), for example, used "boosted SVM" to predict the postoperative life expectancy. These authors have solved the imbalanced data problem towards extracting the decision rules from boosted SVM by applying an "oracle-based" approach. Danjuma (2015) analyzed the performance of MLP vis-à-vis J48 and the NB algorithm on the UCI ML repository dataset for thoracic surgery. From the analysis, MLP was found to perform the best with a classification accuracy of 82.3% vis-à-vis J48 and NB. Kourou, et al. (2015) evaluated predictive models based on various supervised ML techniques such as SVM, ANN, Bayesian networks, and DT with the aim to model cancer risk or patient outcomes. Notably, with Bayesian model, an accuracy of 91.28% was achieved on the same UCI repository dataset from Wroclaw Thoracic Surgery Centre, Poland. To improve on ML techniques when the datasets have a large number of features or attributes, Desuky & El Bakrawy (2016) employed attribute ranking and selection to identify the most relevant attributes while removing those redundant and irrelevant attributes from the dataset. All four (4) of their applied ML algorithms (SVM, LR, MLP, and J48) have also been compared with their boosted versions. Their results showed that boosting is not always the better choice.

In another body of work, Sindhu, et al. (2014) used six (6) classification approaches, including NB, J48, Partial Decision Tree (PART), One R, Decision Stump (DS), and RF to analyze thoracic surgery data. From the analysis, it was found that RF has the best classification accuracy with all split percentages. Nachev & Reapy (2015) studied the chance of patient survival after undergoing post-thoracic surgery by applying data mining techniques for medical diagnosis. Models used included DT, NB, and SVM. Results showed that SVM is the most suited one vis-à-vis other models in term of accuracy.

More recently, with the mushrooming of ML algorithms, the call for better accuracy gained further attention. Kittipat, et al. (2018) have employed the Bayesian network model towards predicting post-thoracic surgery life expectancy as performed on the UCI thoracic surgery dataset. Their experimental results unveiled an accuracy of 91.28% for the Bayesian model with discretization and learning scheme. Zhangheng, et al. (2020) have developed an artificial intelligence (AI) model for

predicting the life expectancy of post-thoracic surgery within a year period. This was done for NSCLC patients with bone metastases by employing the Extreme Gradient Boosting (XGBOOST) algorithm. XGBoost was further compared with SVM, RF, LR towards generating predictive models. XGBoost outperformed other models in terms of accuracy for training and validation with an accuracy of 78.6% being achieved for XGBOOST during validation vis-à-vis other models.

Altogether, ML algorithms clearly have prevailed for predicting post-thoracic surgery life expectancy. Many have even applied a boosted ML version and explored with various decision rules to handle data imbalance. Importantly, where the dataset is high with different attributes and features, attribute ranking and selection are employed, and ML models such as SVM, LR, MLP, J48 and others with boosted version are recommended to improve prediction accuracy (Desuky & El Bakrawy, 2016). To date, results have shown that ANN/MLP and Bayesian model appeared to have achieved the best classification accuracy of about 82.3% and 91.28% respectively (Danjuma, 2015; Kourou, et al., 2015; Kittipat, et al., 2018) Although the use of Bayesian models has often resulted in a higher accuracy, these models are based on directed acyclic graphs, which represent independent (and dependence) relationships between variables. The links in the model represent conditional relationships in the probabilistic sense. Compared to DNN, these models are much simpler with no hidden layers, weights, biases, and activation function for producing the output. Also, there is no concept of back propagation to reduce the gradient loss.

With the current trend towards DNN, a part of deep learning, our focus here is on using DNN to predict the post-thoracic life expectancy of patients with superior accuracy and reduced error, thereby benefitting the healthcare industry. A superior accuracy can be expected because DNN is an advanced form of NN with multiple hidden layers. Thus, results from this work would especially benefit patients and hospital management.

3. THE PROPOSED WORK

Patients undergoing thoracic surgical medical conditions believe that the medication would improve their lifestyle so that they can lead a longer and peaceful life. But it is highly challenging to monitor the survival rate of patients within a year's time post- thoracic surgery. If a pattern exists in the patient dataset pertaining to age, health condition, and other parameters, it would be beneficial in predicting the life expectancy of the patient within a year period. This prediction would be really helpful for the surgeons and the patients in making a more informed decision on whether they should go ahead with performing the surgery or if they would like to pursue palliative care or some other alternative treatments. This information could also be used by clinical researchers to consolidate any useful findings with other research findings to uncover new discoveries.

With the advent of DL, a gap exists as no work has yet been reported in predicting life expectancy of post thoracic surgery patients via DNN. DNN is an advanced form of neural network and a subset of DL. A highly accurate predictive model based on the 17 attributes pertaining to thoracic surgery for life expectancy within a year period can thus be expected with the application of DNN vis-à-vis the more traditional ML techniques . On this basis, we shift focus on the theoretical background behind these machine learning algorithms used for our work been discussed in brief:

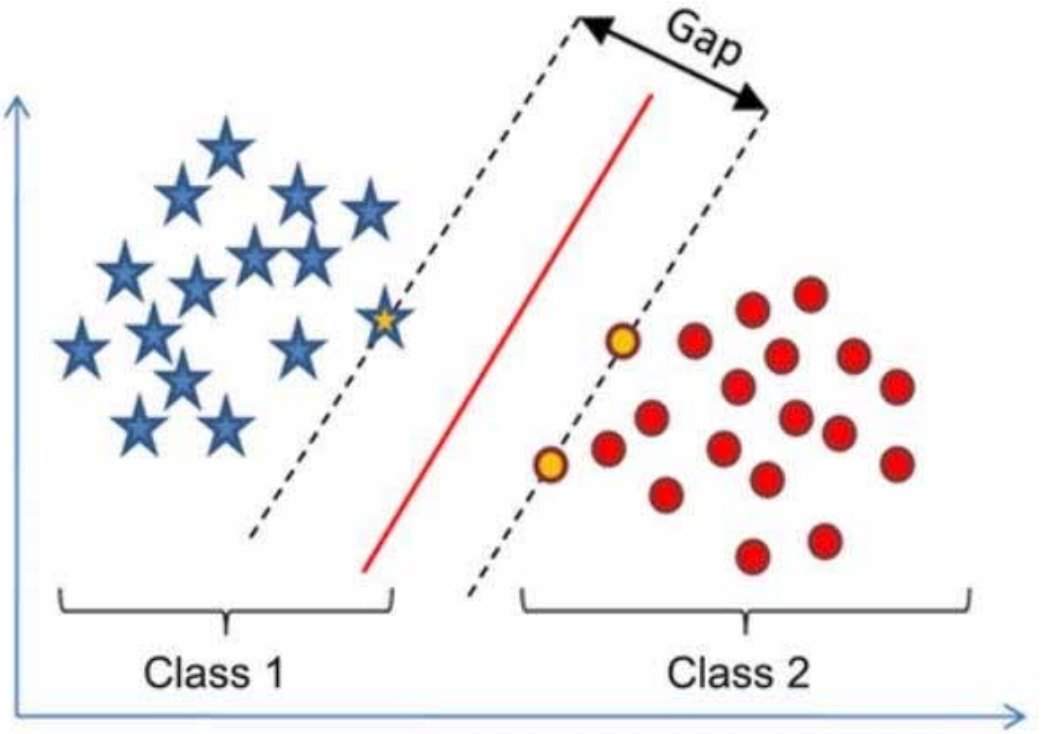
3.1 Popular ML Algorithms

Using traditional ML algorithms such as LR, RF classifier, SVM, KNN, and NB, the thoracic surgery dataset comprises labeled training data, which are categorized as binary classification. Here, we implement a DNN for the life expectancy prediction problem for post thoracic surgery with the following considerations.

3.1.1 Support Vector Machine (SVM)

SVM is used predominantly for classification and regression (Gandhi, 2018; Geron, 2017). **Figure 2** shows the SVM algorithm being plotted as points in space.

Figure 2.



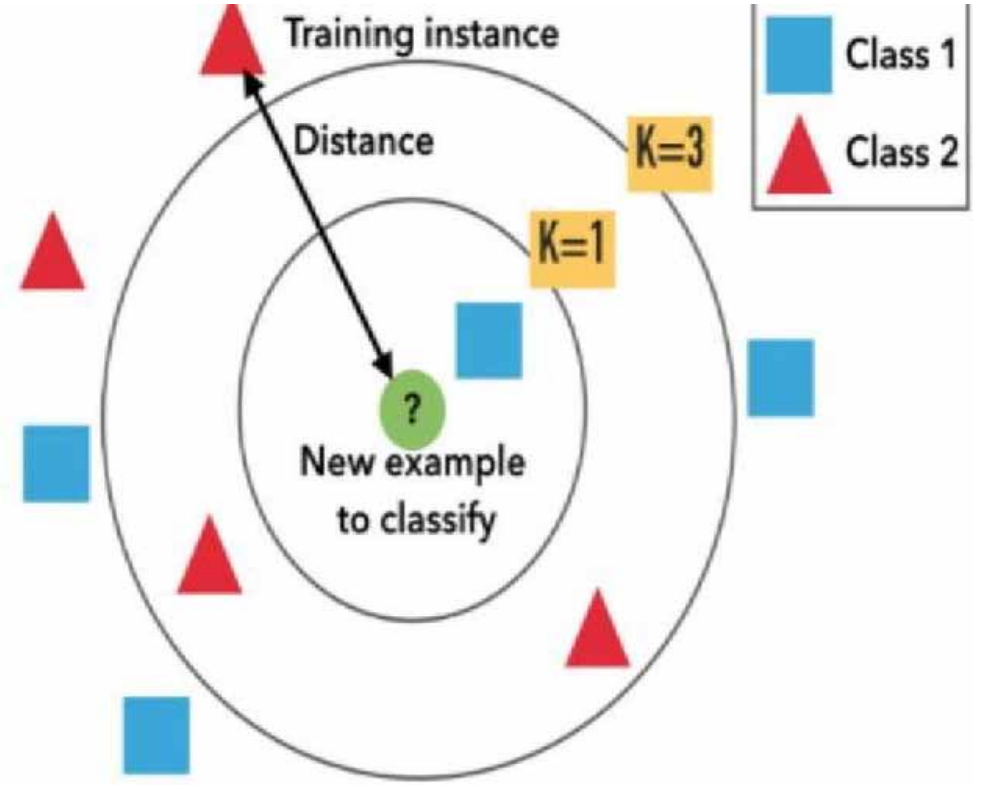
SVM model is generally designated to one class or towards developing a binary or probabilistic linear classifier. In this method, each example belongs to one class or the other which are divided by visible space or gap. A line, known as a *hyperplane*, divides or demarcates the SVM algorithm during classification. A hyperplane is a line that splits the dataset into two halves where each stores the data from two previously established classes. The construction of the hyperplane is carried out in this algorithm where the classification of new values is constructed. After applying the SVM by constructing a hyperplane on a given dataset, data gets classified into different classes. Based on this classification, the prediction accuracy is then computed.

3.1.2 K-Nearest Neighbors (KNN)

As shown in **Figure 3**, KNN is one of the most non-parametric algorithms used for regression and classification (Geron, 2017; Subramanian, 2019).

In nearest neighbors, a particular number of samples that are closer in distance to new points are found, and from that basis, the labels are predicted. The number of neighbors or samples are user defined. In KNN, distance is calculated which can be of any metric and employs the most commonly used Euclidean distance as given in *Equations 1* and *2* below.

Figure 3.



$$d(p,q) = d(q,p)$$

$$= \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2} \quad (1)$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad (2)$$

KNN simply stores the instance of training data only. Classification in this algorithm is performed by computing votes of the nearest neighbor of each point. The output is the average values of its neighbors present around it. The major problem in KNN algorithm is choosing the "K" value. The main drawback is the difficulty in finding the number of nearest neighbors for each sample

3.1.3 Naïve Bayes (NB)

Bayes' theorem is the basis of the NB technique, which assumes that the predictors are independent (Geron, 2017; Brownlee, 2019). It further assumes that there are no features that are related to one

another; simply, the features are completely independent. The mathematical statement of the "Bayes' theorem" is as follows:

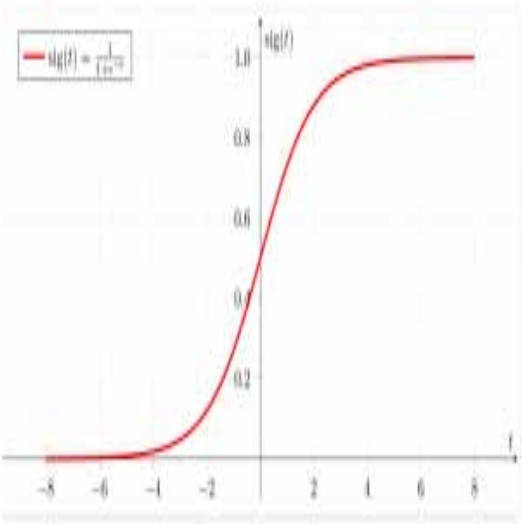
$$P(A \mid B) = P(B \mid A)*P(B)/P(A) \quad (3)$$

Here, P(A) is the prior probability of event A and P(A|B) refers to event A's probability after seeing the evidence. This model is easy to build and may be used with large datasets. The first step in this algorithm is to convert the data set into a frequency table and after creating a likelihood or frequency table, the Naive Bayesian formula is applied to calculate the probability of each class.

3.1.4 Logistic Regression (LR)

LR is a classification algorithm where observations are assigned to a discrete set of classes (Geron, 2017; Swaminathan, 2018). As shown in **Figure 4**, the regression model is built to predict the probability where a given datum belongs to the category number as "1". Only when a "decision threshold" is considered and brought into the picture, LR becomes a classification technique.

Figure 4.



Formally, the LR model may be represented as:

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + x \cdot \beta \quad (4)$$

On solving for p, we get:

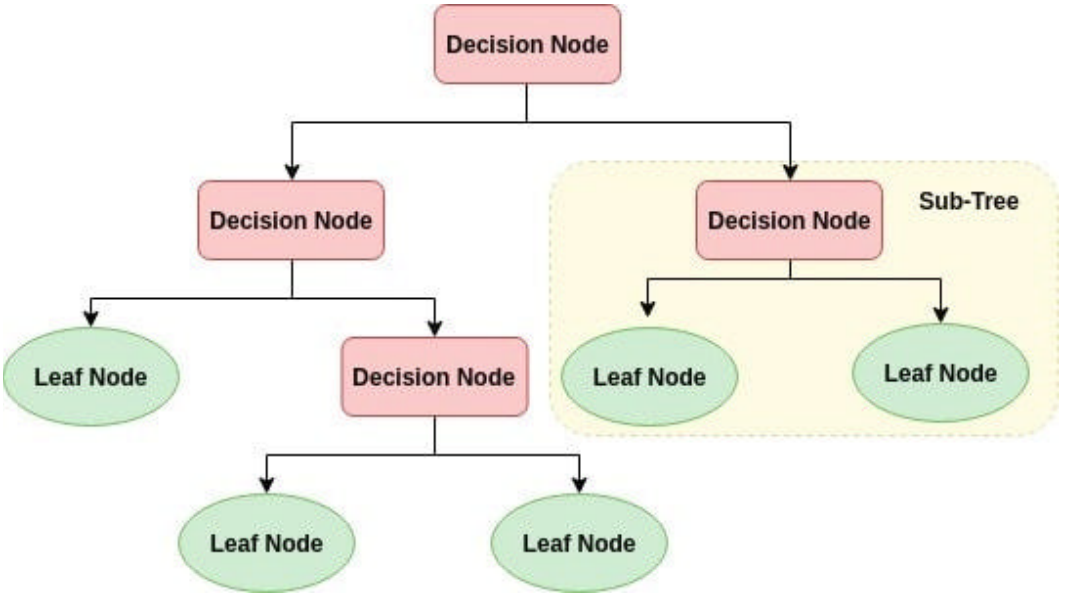
$$p(x;b,w) = \frac{e^{\beta_0 + x \cdot \beta}}{1 + e^{\beta_0 + x \cdot \beta}} = \frac{1}{1 + e^{-\beta_0 + x \cdot \beta}} \quad (5)$$

Fixing the threshold value is the most important one in logistic regression and is dependent on the "classification" problem.

3.1.5 Decision Tree (DT)

A set of the axis-parallel hyperplane dividing the region into a hypercube, the DT is based on the nested if-else classifier (Geron, 2017; Gupta, 2017). It is a classification or regression model in the form of a tree structure being built via the decision trees. As shown in **Figure 5**, DT can handle both categorical as well as numerical data.

Figure 5.



Given that the dataset is broken down into smaller subsets with increasing tree depth, the final result is a tree with decision nodes and leaf nodes. In a DT, the *root node,* which refers to the best predictor, is at the topmost. Two or more branches are created via a decision node whereas the classification (decision) is denoted by the leaf node.

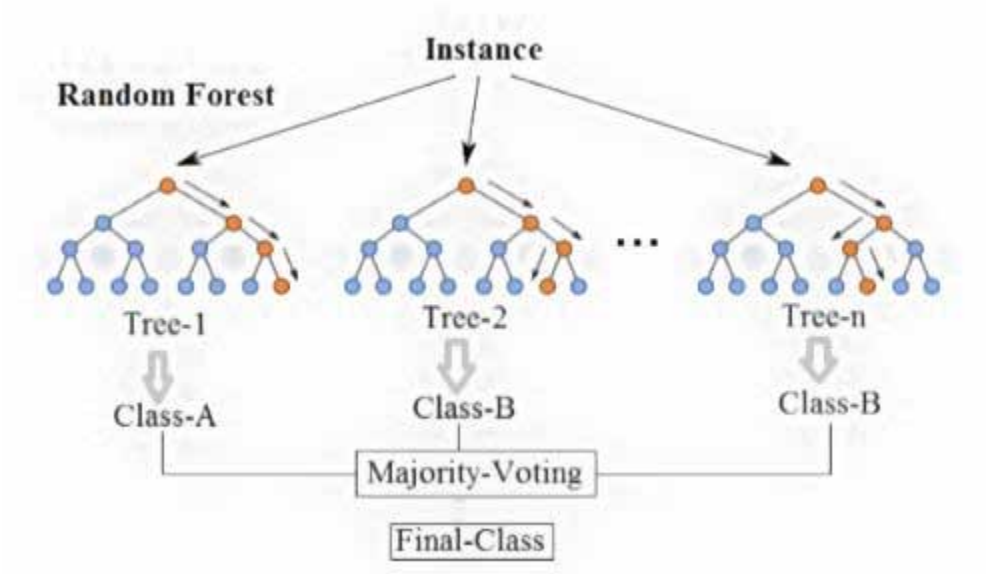
- Construction of DT

Step 1: First, we calculate the entropy of the target
Step 2: Based on different attributes, the dataset is split; following that, entropy is assessed and added proportionally to compute the total entropy for the given split. The resulting entropy assessed is subtracted from the entropy before the split, resulting in information gain
Step 3: Then, we choose the decision node, which divides the dataset by its branches and repeats the same process on every branch.
Step 4a: A branch with an entropy of 0 is noted as a leaf node
Step 4b: A branch with an entropy of more than 0 requires further splitting.

3.1.6 Random Forest (RF)

As shown in **Figure 6**, RF is a classification algorithm comprising multiple decisions trees (Geron, 2017; Yiu, 2019).

Figure 6.



RF uses bagging and feature randomness while building every individual tree to create an uncorrelated forest of trees. The RF technique does both row sampling and column sampling with DT as a base. Owing to column sampling, model h1, h2, h3, h4 are more different than by doing only bagging.

With an increasing number of base learners, variance decreases; also, as the value of k decreases, there will be an increase in the variance. For the entire process, bias remains constant. The value of "k" can be found using the cross-validation technique; in this method, low bias and high variance are needed for our base learner.

- Steps for implementing a RF classifier are as follows:
  1. Consider a training data set of N observations and M features. A sample of data is taken from the training data set randomly with replacement;
  2. Next, the subset of M features is chosen randomly; accordingly, the feature with the best split is used for splitting the node sequentially;
  3. The tree grows as large as possible;
  4. Repeat Steps 1 to 3; further, the prediction is performed in line with the aggregation of predictions from the multiple numbers of trees.
- Train and run-time complexity:

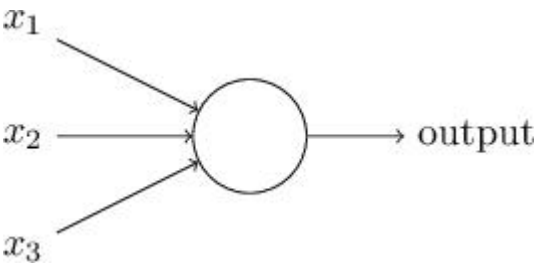
  Training time = O (log(nd)*k)
  Run time = O (depth*k)
  Space = O (store each DT*K)"

As the number of base models increases, training run time increases with an increasing number of the base model. Hence, the use cross-validation to find the optimal hyperparameter is recommended.

3.1.7 Deep Neural Network (DNN)

A perceptron is also known as an artificial neuron forming the neural system (Geron, 2017; Allibhai, 2018). As shown in **Figure 7**, x1, x2, x3 are given as inputs to the perceptron, which produces a single binary output. Algebraically, that is everything as to how a perceptron function.

Figure 7.

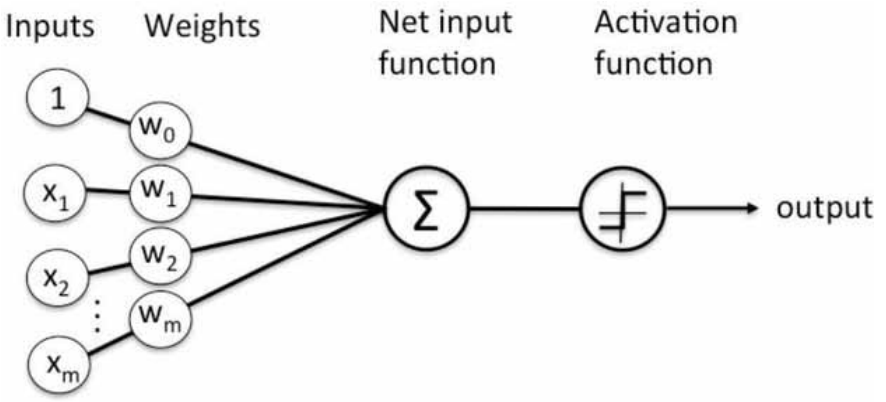


The functioning of the human brain is imitated by employing neural network (NN) technology to uncover pattern recognition rather than passing the input through the different layers of the simulated neural connection.

$$output = \begin{cases} 0 & if \Sigma_j w_j x_j \leq \quad threshold \\ 1 & if \Sigma_j w_j x_j > \quad threshold \end{cases}$$

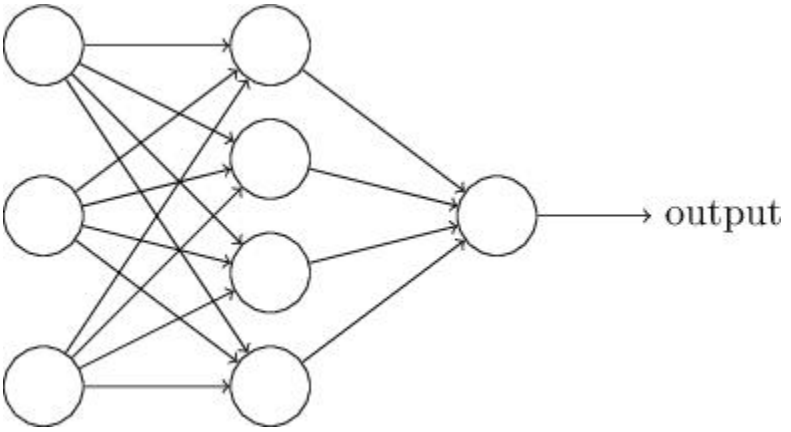
ANN have an input layer, at least one hidden layer in-between and an output layer. In feature hierarchy, specific sorting and ordering types are carried out in each layer. To deal with unlabelled or unstructured data is among the significant uses of these NNs. **Figure 8** shows the Perceptron in ANN.

Figure 8.



Assuming we have the network as shown in **Figure 9**,

Figure 9.



The leftmost layer refers to input neurons present in the input layer whereas the rightmost layer refers to the output neurons present in the output layer. The middle layer refers to the hidden layer, which does not contain the neurons of input or the output. One of the negative or the downside of NN is the cost work slope processing while one of the quicker ways to deal with slope processing is error backpropagation, which gives an in-depth knowledge on altering the metrics towards the system's behavior. Hierarchical composition of linear v. non-linear activation function is given by DNN (Brownlee, 2018).

We use DNN, which is a subset of DL here, comprising an input layer, two hidden layers, and a final output layer. The former layers will be activated via function, ReLu, while the output layer will be activated via function, Sigmoid.

3.2 System Architecture

Figure 10 shows the dataflow of post-thoracic surgery life expectancy system and its interaction with the patient, surgeon, and the hospital system. It shows how the prediction model plays a very crucial role in making a decision towards surgery for patients.

The dataflow diagram depicts a ML-enabled post-thoracic life expectancy prediction system being integrated into the hospital system. Here, the patient makes an appointment for thoracic surgery with the patient's medical data being fed into the ML-based prediction system to forecast the life expectancy within a year after surgery. Based on the analysis, the surgeon advises to perform the surgery or not. The information is then passed on to the patient for treatment, billing, and reports. Notably, the surgeon's recommendation is also passed onto the hospital and stored in the hospital cloud as part of a dataset for further research. This is the key role played by the ML-enabled prediction system in interacting between the patient and surgeon prior to a surgery decision. Also, data collected are stored in the cloud for continuously training and updating the model for better prediction.

Use case diagrams of the system developed is now shown in **Figure 11**.

Four (4) key actors: the nurse, the surgeon, the patient, and the lab technician are involved in this use case. First, the patient interacts with the nurse for a thoracic surgery appointment. Accordingly, the nurse makes the appointment with the surgeon and advises the patient for the respective tests and referral to the assigned surgeon for the consultation. The patient here gives the test samples which are collected by the lab technician. Next, the surgeon examines the patient, analyses the symptoms,

Figure 10.

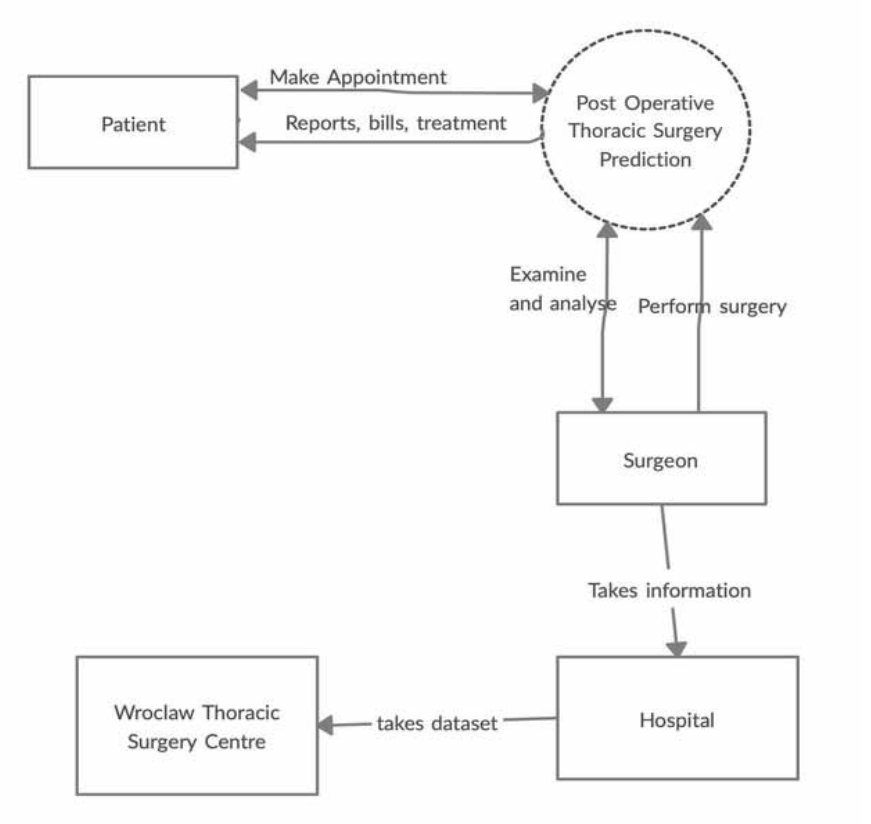
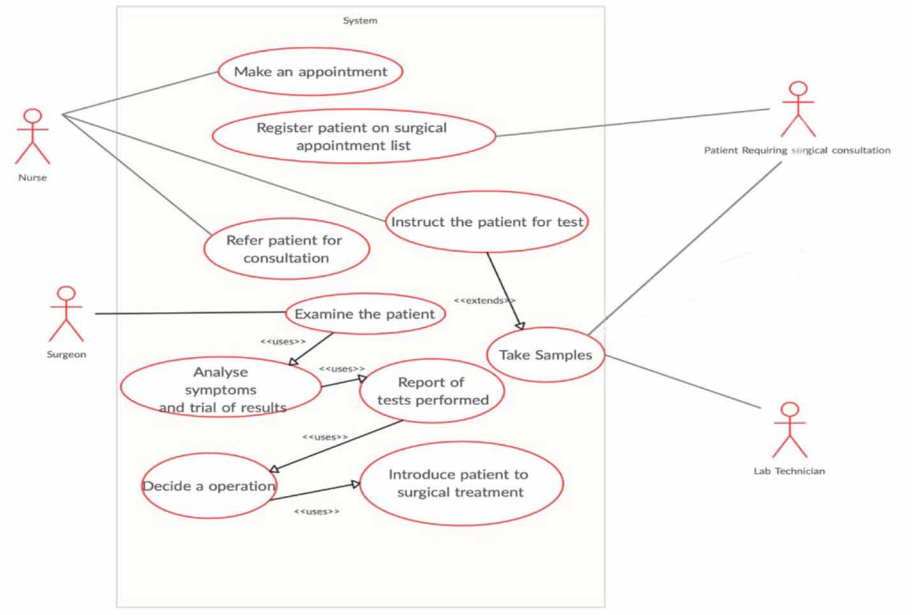


Figure 11.



together with the test results and feeds the relevant information into the prediction system. Based on the output from the prediction system, surgery is ultimately decided, followed by having to prepare the patient for surgery.

In summary, the prediction system becomes an implicit part of the use case diagram between the surgeon and patient in deciding on the surgery based on post-thoracic surgery life expectancy output vis-à-vis the test results and symptoms.

4. RESULTS & ANALYSIS

Plenty of work is involved in determining and acquiring the study dataset. First, hospitals in different countries were consulted on the requirements to perform the lung cancer surgery. As shown in **Figure 12**, retrospective data were drawn from the Wroclaw Thoracic Surgery Centre for patients who had underwent lung surgeries for primary lung cancer between the years 2007 to 2011.

Figure 12.

@data															
DGN2	2.88	2.16	PRZ1	F	F	F	T	T	OC14	F	F	F	T	F	60 F
DGN3	3.4	1.88	PRZ0	F	F	F	F	F	OC12	F	F	F	T	F	51 F
DGN3	2.76	2.08	PRZ1	F	F	F	T	F	OC11	F	F	F	T	F	59 F
DGN3	3.68	3.04	PRZ0	F	F	F	F	F	OC11	F	F	F	F	F	54 F
DGN3	2.44	0.96	PRZ2	F	T	F	T	T	OC11	F	F	F	T	F	73 T
DGN3	2.48	1.88	PRZ1	F	F	F	T	F	OC11	F	F	F	F	F	51 F
DGN3	4.36	3.28	PRZ1	F	F	F	T	F	OC12	T	F	F	T	F	59 T
DGN2	3.19	2.5	PRZ1	F	F	F	T	F	OC11	F	F	T	T	F	66 T
DGN3	3.16	2.64	PRZ2	F	F	F	T	T	OC11	F	F	F	T	F	68 F
DGN3	2.32	2.16	PRZ1	F	F	F	T	F	OC11	F	F	F	T	F	54 F
DGN3	2.56	2.32	PRZ0	F	T	F	T	F	OC12	F	F	F	F	F	60 F
DGN3	4.28	4.44	PRZ1	F	F	F	F	F	OC12	F	F	F	T	F	58 F
DGN3	3	2.36	PRZ1	F	F	F	T	T	OC11	F	F	F	T	F	68 F
DGN2	3.98	3.06	PRZ2	F	F	F	T	T	OC14	F	F	F	T	F	80 T
DGN3	1.96	1.4	PRZ1	F	F	F	T	F	OC11	F	F	F	T	F	77 F
DGN3	4.68	4.16	PRZ1	F	F	F	T	F	OC12	F	F	F	T	F	62 F
DGN2	2.21	1.88	PRZ0	F	T	F	F	F	OC12	F	F	F	T	F	56 F
DGN2	2.96	1.67	PRZ0	F	F	F	F	F	OC12	F	F	F	T	F	61 F
DGN3	2.6	1.68	PRZ1	F	F	F	T	F	OC12	F	F	F	T	F	70 F
DGN3	2.88	2.48	PRZ0	F	F	F	F	F	OC11	F	F	F	T	F	71 F
DGN3	4.48	3.48	PRZ0	F	F	F	F	F	OC12	F	F	F	T	F	51 F
DGN4	3.32	2.84	PRZ0	F	F	F	F	F	OC12	F	F	F	T	F	62 F
DGN3	2.36	1.68	PRZ0	F	F	F	F	F	OC12	F	F	F	T	F	62 F
DGN3	3.68	2.32	PRZ0	F	F	F	F	F	OC11	F	F	F	T	F	62 F
DGN8	4.32	3.2	PRZ0	F	F	F	F	F	OC11	F	F	F	F	F	58 T
DGN5	4.56	72.8	PRZ0	T	T	F	T	F	OC12	F	F	F	T	F	57 F
DGN3	3.24	3.08	PRZ1	F	F	F	T	F	OC11	F	F	F	T	F	60 F
DGN3	3.4	3.06	PRZ1	F	F	F	T	T	OC11	F	F	F	T	F	68 T
DGN3	3.16	2.69	PRZ1	F	F	F	T	T	OC11	F	F	F	T	F	56 F
DGN6	3.96	3.28	PRZ0	F	F	F	F	F	OC11	F	F	F	T	F	61 F

This centre is associated with the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland. The research database constitutes a part of the National Lung Cancer Registry, administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland (Lubicz, et al., 2013).

The data are presented in the form of rows containing the patients (470 training dataset) and columns comprising features; specifically, sixteen (16) features with true-false labelling were used in developing a ML model for prediction. These features are key to predicting life expectancy of post-thoracic surgery for patients having had the surgery.

Examples are labelled on the basis of whether the given patient ultimately lived or died. A “false” label specifies that the patient lived 1 year after the surgery, while a “true” label specifies that the patient died within a year after the surgery. Features included continuous data and class data on the patients. Some of the continuous data comprise the patient’s age at the time of surgery together with factors such as the size of the original cancerous tumor(s), the respective patient smoking history and past asthmatic problems as well as the maximum respective volume that the lungs would exhaled.

Other features included were signs of coughing prior to surgery, the presence of pain and haemoptysis before surgery and whether the patient was a smoker or has had asthma, among other items. This classification is further used to predict whether the patient survived the one-year period or not. Three scale variables: age, volume, and capacity were noted. Additionally, the dataset observations included nominal variables such as diagnosis-specific combination of ICD-10 codes for primary v. secondary as well as multiple tumors, if any as noted below.

- DGN: Diagnosis - specific combination of ICD-10 codes for primary and secondary as well as multiple tumor(s), if any (DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1)
- PRE4: Forced vital capacity - FVC (numeric)
- PRE5: Volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric)
- PRE6: Performance status - Zubrod scale (PRZ2, PRZ1, PRZ0)
- PRE7: Pain before surgery (T, F)
- PRE8: Haemoptysis before surgery (T, F)
- PRE9: Dyspnoea before surgery (T, F)
- PRE10: Cough before surgery (T, F)
- PRE11: Weakness before surgery (T, F)
- PRE14: T in clinical TNM - the size of the original tumor, from OC11 (smallest) to OC14 (largest) (OC11, OC14, OC12, OC13)
- PRE17: Type 2 DM - diabetes mellitus (T, F)
- PRE19: MI up to 6 months (T, F)
- PRE25: PAD - peripheral arterial diseases (T, F)
- PRE30: Smoking (T, F)
- PRE32: Asthma (T, F)
- AGE: Age at surgery (numeric)
- Risk1Y: 1 year survival period - (T)rue value if died (T, F)

As ML algorithms work better with integer and floating values rather than string values, the dataset was then modified as detailed in **Figure 13**

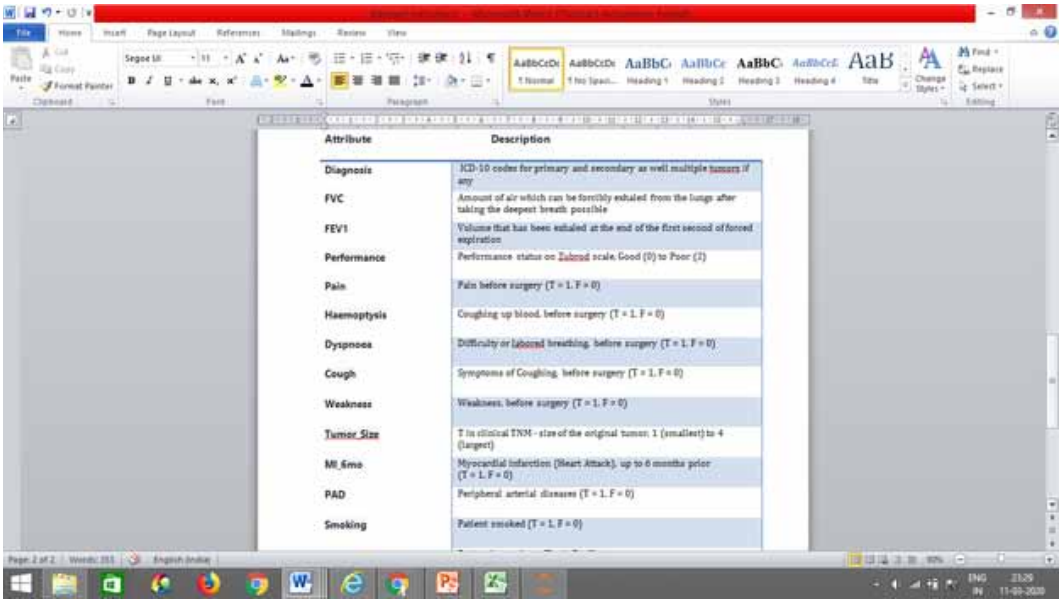
After analyzing the unlabeled information in the dataset, many columns were observed to contain string type data for true/false values. Hence, in order to reduce the redundancy and improve on the accuracy for better analysis, we converted the string objects into integer and float values with T/F objects into 1 v. 0 integer data types. The columns are renamed DGN to human-readable format and labeled with various attributes such as FVC, FEV1, smoking, asthma, and more.

3.2 Metrics

Results of the algorithms are compared to identify the best algorithm for the prediction of the post-thoracic surgery life expectancy. Brief explanation on the relevant metrics studied are given below.

- Confusion Matrix: This represents a binary classifier where different parameters were fed into the DNN system
- TP v. TN refer to the correctly classified instances as *lived* v. *dead* occurrences
- FP v. FN refer to the wrongly classified instances as *no death* v. *lived*
- Accuracy: This is computed with the help of the below formula

Figure 13.



$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- Precision: This refers to the positive predictive value, which is computed as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall: Also called sensitivity, which is computed as:

$$\text{Re}\,call = \frac{TP}{TP + FN}$$

- F1 Score: This metric takes into account the recall and precision values and is computed as:

$$F1score = 2 * \frac{\text{Pr}\,escision * \text{Re}\,call}{\text{Pr}\,ecision + \text{Re}\,call}$$

4.2 Performance of ML Algorithms

The dataset was split into training v. testing samples at 70:30. ML algorithms, including LR SVM, NB, KNN, DT, RF, and DNN, were evaluated on the training dataset. After the dataset has been

trained via these algorithms, the performance of the various ML algorithms in terms of precision, recall, F1 score and accuracy was validated via the testing dataset with results as tabulated in **Table 1**.

Table 1 shows the DNN outperforms other ML models in terms of accuracy, achieving an accuracy of 91.4% vis-à-vis other models. Comparative analysis in terms of accuracy has been plotted as a graph in **Figure 14**, validating the reported results.

Table 1. Performance Metrics of Machine Learning Model

Machine Learning Model	Precision	Recall	F1 Score	Accuracy
Logistic Regression	0.72	0.85	0.78	85%
Support Vector Machine	0.72	0.85	0.78	85%
Naïve Bayes	0.74	0.73	0.73	73%
K-Nearest Neighbor	0.74	0.73	0.73	73%
Decision Tree	0.74	0.73	0.73	73%
Random Forest	0.72	0.83	0.77	83%
Deep Neural Network	0.87	0.50	0.63	91.4%

Figure 14.

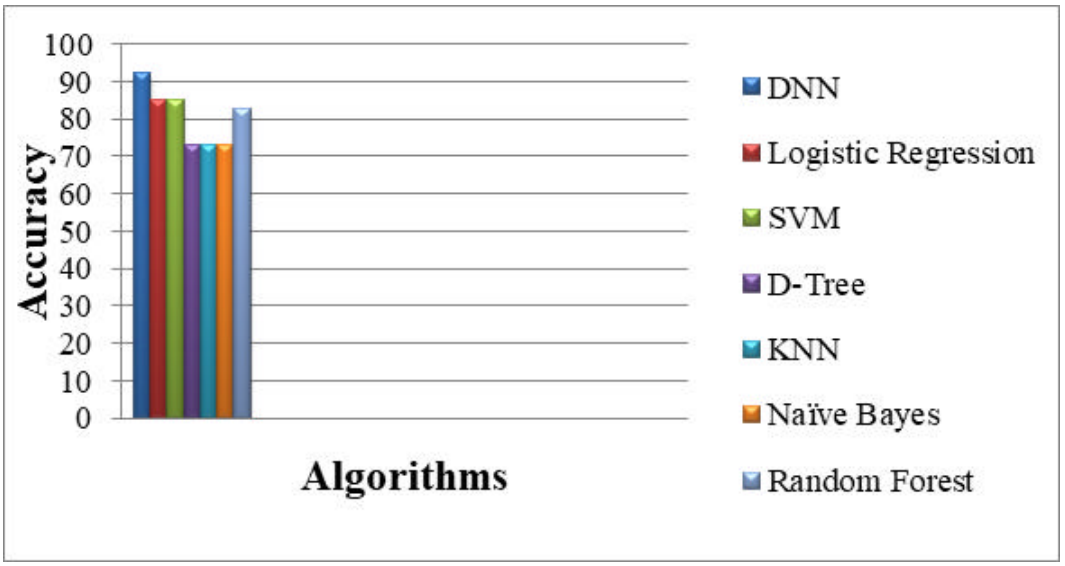
Table-1: Performance Metrics of Machine Learning Model

Machine Learning Model	Precision	Recall	F1 Score	Accuracy
Logistic Regression	0.72	0.85	0.78	85%
Support Vector Machine	0.72	0.85	0.78	85%
Naïve Bayes	0.74	0.73	0.73	73%
K-Nearest Neighbor	0.74	0.73	0.73	73%
Decision Tree	0.74	0.73	0.73	73%
Random Forest	0.72	0.83	0.77	83%
Deep Neural Network	0.87	0.50	0.63	91.4%

As well, **Figure 15** depicts the ROC curve for DNN, which plots between true positive rate (sensitivity) and the false positive rate for the different cut-off points of a parameter for the DNN.

In **Figure 15**, the ROC curve is around 0.9, nearing the maximum value of 1.0, showing the best performance. For *precision*, it indicates what proportion of positive prediction was correct. In DNN, this means that 87% of times, the system has predicted the life expectancy as correct on whether the patient would be alive or not. For *recall*, the system indicates what proportion of actual positive was identified correctly. For the DNN model, 50% of the time the system has correctly identified the life expectancy of patients as alive or not correctly.

Figure 15.



As for F1 score, a metric which takes both precision and recall, we achieved a score of 0.63 for the DNN model. This score is acceptable as there is no negativity in the model. The reason for the lower value in DNN for recall and F1 score vis-à-vis other models could be due to the smaller testing dataset. DL always requires a massive data set, a key limitation of the current study.

With a larger data set, the DNN model will be able to achieve a better recall and F1 scores, with even higher accuracy.

5. CONCLUSION

For lung cancer patients, irrespective of the cancer being one of SCLC or NSCLC, the survival rate post-surgery is relatively low. Hence, a thorough diagnosis and analysis based on past patient historic data and the medical condition of the patient are needed prior to recommending surgery. Other considerations include factors such as age, experience of the surgeon, and more.

Prior research has employed ML in post-thoracic surgery life expectancy prediction for patients with lung cancer. Many ML models such as MLP, SVM, NB, DT, LR and RF have already been applied on relevant datasets to predict post-thoracic surgery life expectancy. Moreover, work on attribute ranking and selection to achieve better prediction accuracy with ML algorithms has also

been conducted. Notwithstanding, a key challenge in past ML algorithms employed is the focus on traditional methods with a gap of no work being reported in the DL-based model applications.

In the current work, the qualities of sixteen (16) attributes and selection methods have been tested to improve the prediction for the life expectancy of lung cancer patients post-thoracic surgery within the one-year period. Results indicate that the DNN can be efficiently used for life expectancy prediction, which provides the best accuracy vis-à-vis other ML algorithms. The evaluation effort confirms that the DNN not only provided better accuracy at 91.4% vis-à-vis other more traditional ML algorithms, but its employment also reduced costs while increasing process efficiencies.

Many aspects of future works can be extended from the current research. First, we could have improved our results by implementing a larger dataset. From the analysis of this dataset, we see that a larger dataset would have yielded better performance in terms of various studied metric as it improves the scope of the model. Hence, the current study can be expanded with larger datasets and more attributes in DNN model in future though the values achieved are currently acceptable for the dataset being studied.

Second, recurrent NN (RNN) is designed to recognize sequential attributes and patterns in the dataset to predict the next most likely scenario. Again, the desired outcome would depend on the hospital or patients and how they view these predictions for life v. death outcomes, and how they the efficiency of the model is to be determined. Future work can evaluate RNN v. DNN alongside other ML algorithms.

Finally, the conditions affecting the post-operative life span have to be taken as a separate research work by developing a mobile application for interacting with the patients while developing an intelligent (even self-automated) recommendation system for guiding the patients to lead a less disruptive but enjoyable life while adopting a more healthy lifestyle for the years they may still have on hand.

REFERENCES

Adam, A., Ivaylo, B., & Peng, J. (2014). *Life Expectancy Post Thoracic Surgery*. Retrieved from http://cs229.stanford.edu/proj2014/Adam%20Abdulhamid,%20Ivaylo%20Bahtchevanov,%20Peng%20Jia,Life%20Expectancy%20Post%20Thoracic%20Surgery.pdf

Allibhai, E. (2018). *Building A Deep Learning Model using Keras*. Retrieved from https://towardsdatascience.com/building-a-deep-learning-model-using-keras-1548ca149d37

American Medical Association. (n.d.). *Thoracic Surgery Specialty Description*. Retrieved from https://www.ama-assn.org/specialty/thoracic-surgery-specialty-description

Brownlee, J. (2018). *A Gentle Introduction to Dropout for Regularizing Deep Neural Networks*. Retrieved from https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/

Brownlee, J. (2019). *A Gentle Introduction to Bayes Theorem for Machine Learning*. Retrieved from https://machinelearningmastery.com/bayes-theorem-for-machine-learning/

Danjuma, K. (2015). Performance Evaluation of Machine Learning Algorithms. *International Journal of Computer Science Issues*, *12*(2), 189–199.

Desuky, A., & El Bakrawy, L. (2016). Improved prediction of post-operative life expectancy after Thoracic Surgery. *Advances in Systems Science and Applications*, *16*(2), 70–80.

Ferlay, J., Shin, H.-R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010). Estimates of Worldwide Burden of Cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer*, *127*(12), 2893–2917. doi:10.1002/ijc.25516 PMID:21351269

Gandhi, R. (2018). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Retrieved from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Geron, A. (2017). *Hands–On Machine Learning with Scikit–Learn and TensorFlow*. O'Reilly.

Gupta, P. (2017). *Decision Trees in Machine Learning*. Retrieved from https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

Kittipat, S., Kittisak, K., & Nittaya, K. (2018). Post-Operative Life Expectancy of Lung Cancer Patients Predicted By Bayesian Network Model. *International Journal of Machine Learning and Computing*, *8*(3), 280–285. doi:10.18178/ijmlc.2018.8.3.700

Kokulu, M., Kahramanli, H., & Allahverdi, N. (2015). Applications of Rule Based Classification Techniques for Thoracic Surgery. *2015 Join International Conference on Technology, Innovation and Industrial Management (TIIM)*, 1991-1998.

Kourou, K., Exarchos, T., Exarchos, K. V., Karamouzis, M., & Fotiadis, D. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17. doi:10.1016/j.csbj.2014.11.005 PMID:25750696

Lubicz, M., Pawelczyk, K., Rzechonek, A., & Kolodziej, J. (2013). *UCI Machine Learning Repository: Thoracic Surgery Data Data Set*. Retrieved from https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data

Nachev, A., & Reapy, T. (2015). Predictive Models for Post-Operative Life Expectancy after Thoracic Surgery. *Mathematical and Software Engineering*, *1*(1), 1–5.

Sarna, L., Cooley, M., Brown, J., Chernecky, C., & Kotlerman, J. (2008). Symptom severity one to four months post-thoracotomy for lung cancer. *American Journal of Critical Care*, *17*(5), 455–468. doi:10.4037/ajcc2008.17.5.455 PMID:18776002

Siegel, , & Miller, , & Jemal. (2020). Cancer statistics. *CA: a Cancer Journal for Clinicians*, *59*(4), 225–249. PMID:19474385

Sindhu, V., Sathya Prabha, S. A., Veni, S., & Hemalatha, M. (2014). Thoracic Surgery Analysis Using Data Mining Techniques. *International Journal of Computer Technology and Applications*, *5*(2), 578–586.

Subramanian, D. (2019). *A Simple Introduction to K-Nearest Neighbors Algorithm*. Retrieved from https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e

Swaminathan, S. (2018). *Logistic Regression — Detailed Overview*. Retrieved from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Timmerman, J., Tönis, T., Dekker-van Weering, M., Stuiver, M., Wouters, M., Van Harten, W., & Vollenbroek-Hutten, M. (2016). Co-creation of an ICT-supported cancer rehabilitation application for resected lung cancer survivors: Design and evaluation. *BMC Health Services Research*, *16*(1), 1–11. doi:10.1186/s12913-016-1385-7 PMID:27121869

Yiu, T. (2019). *Understanding Random Forest*. Retrieved from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Zhangheng, H., & Chuan, H., Changxing, C., Zhe, J., Yuxein, T., & Chengliang, Z. (2020). An Artificial Intelligence Model for Predicting 1-Year Survival of Bone Metases in Non-Small Cell Lung Cancer Patients Based on XGBOOST Algorithm. *Hindawi Biomedical Research International*, *2020*, 1–13.

Zieba, M., Tomczak, J., Lubicz, M., & Swiatek, J. (2014). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing*, *14*, 99–108. doi:10.1016/j.asoc.2013.07.016

Akshaya Ravichandran is currently working as Associate Software Engineering in EY GDS Bangalore. She holds B.tech Degree in Information technology in 2020 from SRM Institute of Science and Technology. Her current interest lies in Machine learning, deep learning. She got one paper published pertaining to Anomaly detection using Deep learning in Lecture Notes in Electrical Engineering.

Krutitka Mahulikar is currently working as Business analyst in Uniphore Software Systems. She holds B.Tech Degree in Information technology in 2020 from SRM Institute of Science and Technology. Her current interest lies in Machine learning, deep learning.

Shreya Agarwal is pursuing her graduate programme in Data science in Maynooth University, Ireland. She holds B.tech Degree in Information technology in 2020 from SRM Institute of Science and Technology. Her current interest lies in Machine learning, deep learning.

Suresh. Sankaranarayanan, PhD., is currently a Full Professor in School of Computing, SRM Institute of Science and Technology, Chennai, India. He has supervised 3 PhDs in Internet of things, AI, Edge Computing. Also he is currently supervising 5 more PhDs in IoT, AI, Edge computing, Block Chain and Medical Imaging. His current research interests are mainly towards 'Internet of Things, Fog Computing, Intelligent Agents, Wireless Networking, Machine Learning". He got more than 100 publications to his credits in major refereed International Journals, Book chapter and Conferences with Google Scholar citation of 1189 with H-index of 19 and Scopus Citation of 506 with H-index of 11.