Breast Cancer Detection using Machine Learning Technique

Vranda Verma, Vidhi Jain,Vaibhav Wadhwa,Vansh Rana,Sheetal Tomar, Dr. Vikas Shrivastava
Department Of Computer Science And Engineering
Meerut Institute Of Engineering And Technology , Meerut

vranda.verma.cs.2019@miet.ac.in , vidhi.jain.cs.2019@miet.ac.in
vaibhav.wadhwa.cs.2019@miet.ac.in , vansh.rana.cs.2019@miet.ac.in ,
sheetal.tomar.cs.2019@miet.ac.in
DOI:10.48047/ecb/2023.12.si4.729

Abstract: According to formal studies, it has been concluded that Breast Cancer is one of the most common cancers which is diagnosed particularly in women. As, Breast Cancer can occur both in men and women but it is mainly found diagnosed in women. Around 14.725% of cases of Breast Cancer are found in India. Many pieces of research are conducted from time to time for early detection of Breast cancer which can help in the timely diagnosis of the disease and the patient can get possible treatment sooner. But It has been found that only 86% of them are diagnosed correctly. Biopsy used to detect cancer contains several shortcomings such as it carries a small risk of bleeding or infection or may cause false defection of disease Hence, the need for a new alternatives method arises that is easy to implement, less risky, reliable, safer, cost more effective and can predict more accurate results. The model proposed in the paper is a combination of the Decision tree, Artificial Neural network K-nearest neighbor & Support Vector Machine.

KEYWORDS: Biopsy, Cancer, Detection, Women

1. Introduction

1.1 General Overview Of The Topic

According to an exploration conducted by the World Health Organization, it is one of the most well-known kinds of disease in ladies. Breast Cancer is also the most common cancer among females in India and is one of the leading causes of death. About 5%.0f Indian women are affected by Breast cancer, while it influences around 12.5% of absolute ladies in Europe and the US. Ladies with Breast Cancer growth in Malaysia are bound to be available at a later phase of the sickness than ladies in different nations.In most cases, Breast Cancer can be easily diagnosed if relevant symptoms occur in the body. On a contrary, it may happen that no symptoms appeared that conclude that some women with breast cancer do not have any symptoms. Hence, BC

screening has taken a major part in the earlier detection of cancerous cells i.e. malignant ones in the patient's body [1][2][3].

Early Recognition of this disease advantages early treatment and determination as it is so basic for long haul endurance. This is due to the reason that if a disease is diagnosed earlier, there is a higher chance for its treatment. Also, the chance of dying from such a disease tends to be reduced as it gets detected earlier and engaged as the main part of a patient's endurance[1][2]. Deferring in recognizing malignant growth at a later stage will prompt the spread of illness in the patient's whole body.A late cancer diagnosis is associated with the disease progressing to its most advanced stages, which reduces the likelihood that the patient will be saved.According to a study conducted by 87 doctors, patients with breast cancer who start treatment within 90 days after the side effects start showing are more likely to survive than those who wait more than that [3][4].From various studies and research across the world, it has been concluded that early detection of breast cancer in a patient's body not only provides early treatment and cure to the disease but also prevents malignant i.e. cancerous cell from spreading all through the patient's body.The paper's primary commitment is an assessment and investigation of the job of different AI approaches in breast cancer early discovery [4][5].

1.2 Literature Review

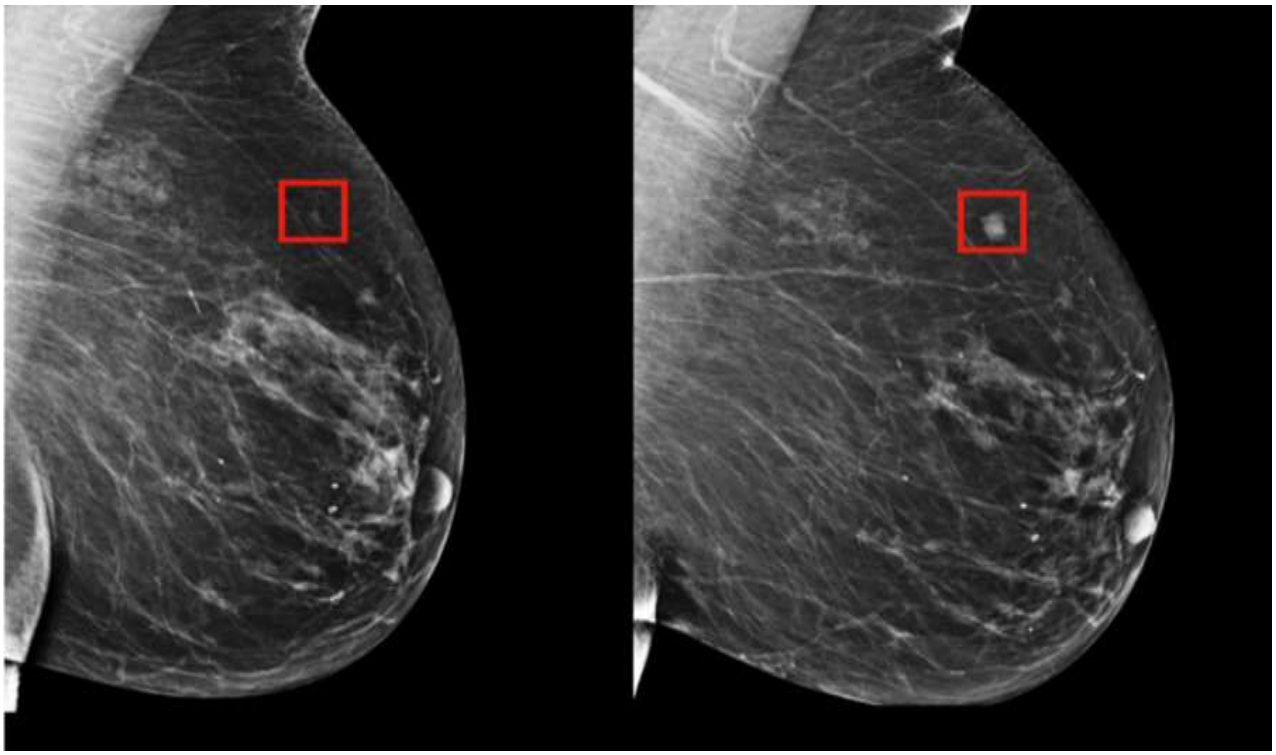
Using Machine Learning methods submerged with each other to predict and diagnose breast cancer not only to prevent excessive treatment for the disease but also to provide more accuracy and integrity to early diagnosis of a disease which helps the patient to provide timely treatment and help doctors and patients to make a more decision to cure and prevent the spread of disease in patient's entire body [6][7][8].

For example, The choice on whether the patient necessities medical procedure depends on the aftereffects of the breast cancer test. Mammograms can give bogus positive outcomes which can prompt pointless techniques. Sometimes benign cells can be found when surgery is performed to remove cancer cells. The patient will be exposed to superfluous, unsavoury, and expensive medical procedure. Medical services-related datasets including images, x-rays, and blood tests can benefit from AI algorithms. While some systems are more applicable for small informational indexes, the rest are better suited for huge informational indexes.From various Studies , it has been profound that the treatment on BC is directly proportional to the Stage of Cancer in which the patient is in. Several symptoms that might give the hint of rise in malignant tissues in the body are Red-Brown(Bloodish appearance) discharge, Absurd change in the shape and size, or might be the uneasiness in the breast[3][5]. The use of Mammograms makes it easier to identify and detect whether the person is suffering from breast Cancer or not ,as early as possible.It has been found that the risk of breast cancer mainly depends on the reproductive history of the patient, the genetic factor , genes mutation [3][4].

Treatment to the patient is decided on the basis of factors or symptoms they exhibit, and based on the physician prescriptions. If the patient is suggested to undergo surgery they might include Tissue expansion, Mastectomy, Lumpectomy,Mammaplasty,Lymph node dissection[1][5].In Medical procedure they are suggested to undergo radiotherapy and teletherapy.The argue to detect the growth of malignant tissue in a person as early as possible give rise to the introduction of most pronic method which both economically and technically feasible .The Detection of Absurd

Change in size and shape of breast using Artificial Intelligence tools is one of those techniques which makes it possible to start the treatment [9][10].

Fig.1 Depiction of Malignant Tissue in breast using AI



2. Proposed Methodology

We will use various Machine Learning Techniques on the real time data collected in a .csv file to figure out the best approach to find the reliable solution which is most significant in almost all aspects.We will try out different techniques such as Support Vector Machine, Boosting , Random Forest Classifier, Bagging, Ensemble method, k-Nearest Neighbour,Decision Tree etc [11][12].

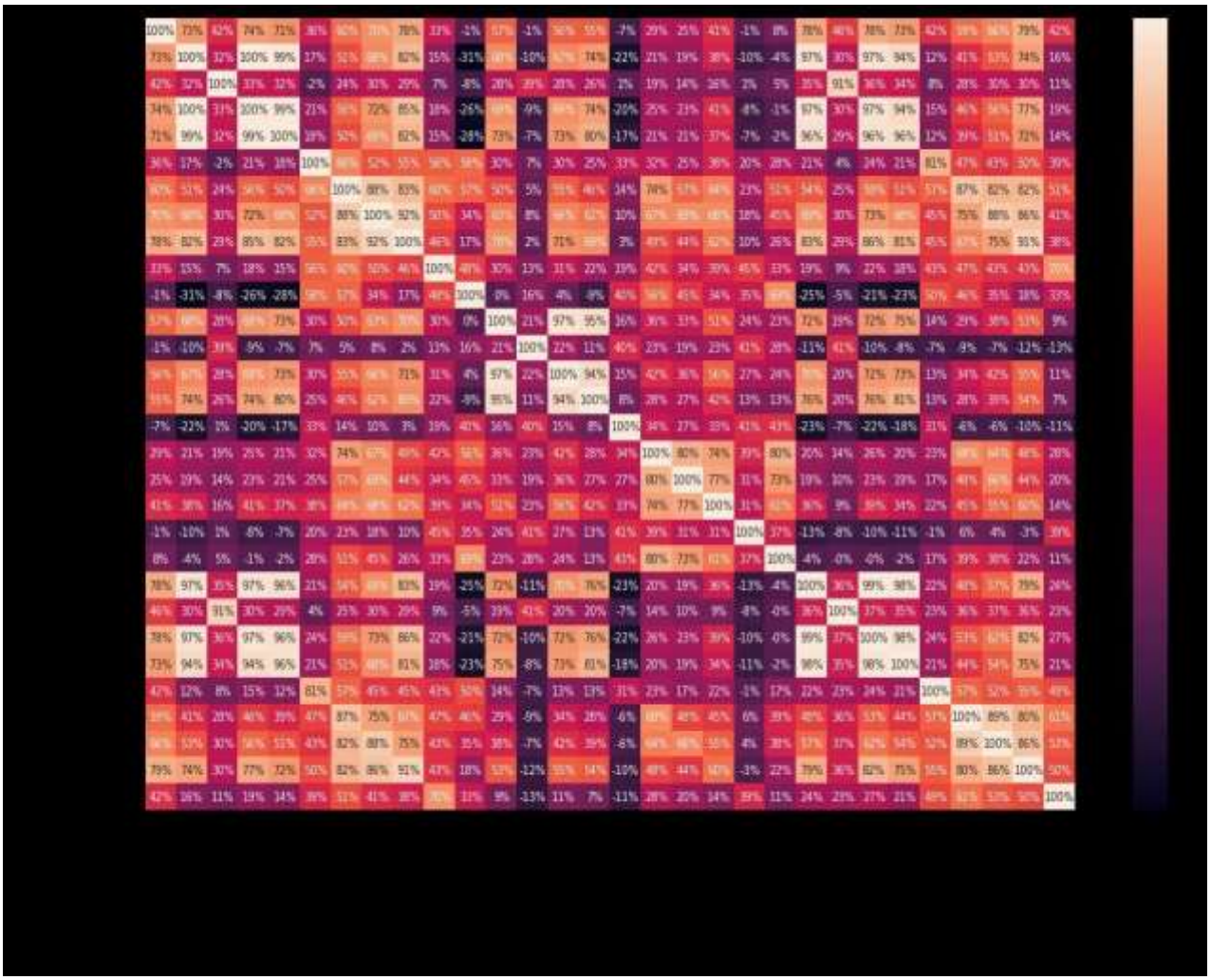
2.1 Boosting

It is one of the techniques used in machine learning so that the error can be minimized in predictive data analysis such as in breast cancer detection. It is used to create a strong classifier over a weak classifier. It simply combines the weak classifier to build a stronger classifier which can help to reduce more training errors as possible. It helps us to transform the weak classifier. It simply combines the weak classifier to build a stronger classifier which can help to reduce more training errors as possible. It helps us to transform the weaker into stronger ones. It helps to predict more accurately whether the person is suffering from Breast cancer or not [13][14][15].

We collect a gathering of models which are amassed towards securing a few strong understudies whose lead is predominant. Each weak model in this collection is fitted, increasing the dependability of the dataset that was before absent.. Every most recent mode highlights one's endeavors for problematic insights reasonable till the ongoing second, therefore the user acquires, close to the completion of the collaboration, a strong understudy with a lower tendency. Helping can be utilized for backslide as well as portrayal issues. Models with modest changes but strong propensity are the base models that are frequently taken into consideration for assistance.

AdaBoost and inclination supporting are the two meta-calculations employed here. Distinct from how they make them, the two strategies seek to take the students in different ways [16][17].

Fig.2 Real Time data describing the percentage of BC tissue



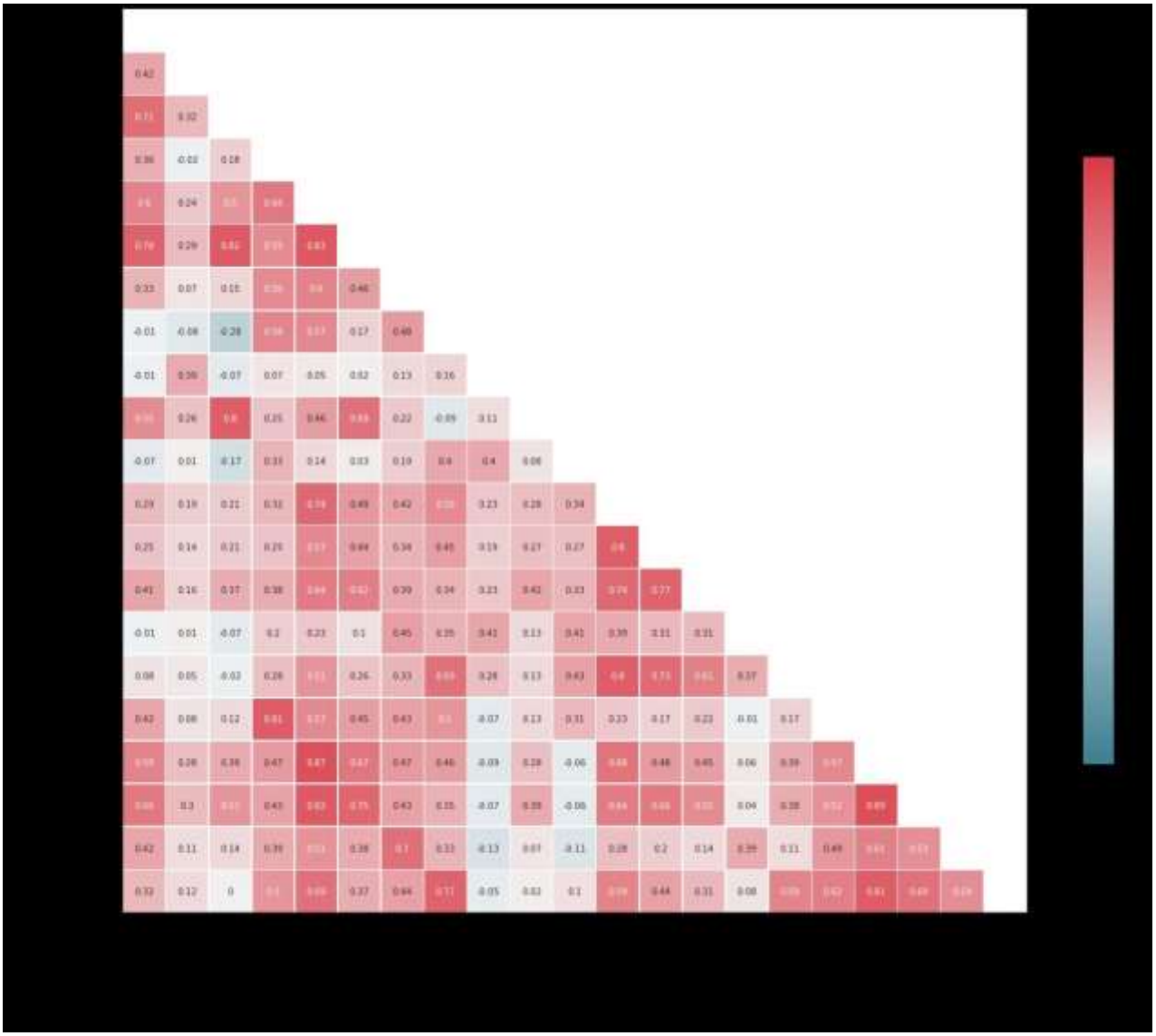
2.2. Support Vector Machine:

Support Classification, regression, and outliers identification all require vector machines. The calculation's aim is ascertain hyper-plane in an M-multilayer space to discriminate between data points. This helps us to provide higher-speed computation and higher performance with a limited number of dataset samples. It is an extremely strong and flexible AI model, fit for performing straight or nonlinear grouping. In sum, the data is taken from the provided dataset and the model is then trained to predict the accurate result. It helps us to analyze whether the cell in the breast is cancerous or not, which helps to detect that the patient is suffering from breast cancer disease.There are countless potential hyperplanes that might be used [18][19][20].

The objective is to find the plane with the most noticeable edge, or the most noticeable gap between the statistics sections of the two classes. In order to group the statistics focal points more reliably, the border interval is enlarged [21][22].

Extending the edge between the hyperplane and the relevant information is the goal.

Fig.3 Vector points that surround hyperplane



2.3. Random Forest Classifier:

It is one of the supervised Algorithms used in Machine Learning. It works by making Decision trees on different data inputs or samples. It is one of the most popular supervised algorithms used in ML. As the name suggests, it is a forest that is developed by the combination of various trees these trees here are the decision tree and we can compute the result by taking the mean of all the trees to predict a more accurate result. It makes decisions based on the number of majority votes available. It has been also seen that in many cases decision trees indigently, are unable to find correct output but collectively many trees help us to predict a more accurate and correct result. It is more reliable as the training need to train the model is lower as compared to other models. It has been noticed that it even works efficiently if the data is missing in the dataset. It is a combination of decision trees. A decision tree uses either Recursive Partitioning or a Conditional Inference Tree for its development. In recursive partitioning, the decision tree is developed by patting nodes. The source set into subsets is what advanced the tree. The recursion stops if a subset's value is close to that of the goal variable. A non-parametric test is included in the contingent inference tree to prevent overfitting. Despite the fact that large trees may use a lot of memory, random forests

can manage lacking qualities, tenacious, out, and resemble data. The HYPER-Boundaries need to be tuned [23][24][25].

The model has the most involvement from Random Forest. It increases effectiveness by using arbitrary examples. Every tree group votes in the election. It is used in independent mode for area inspection. The bootstrap test is used to fit trees as part of the random forest strategy, which combines the results. Another method used by random forests to make the various fitting trees less related to one another is to merely test over the main part of the random subset while cultivating each tree rather than only looking at the examples' conclusions.

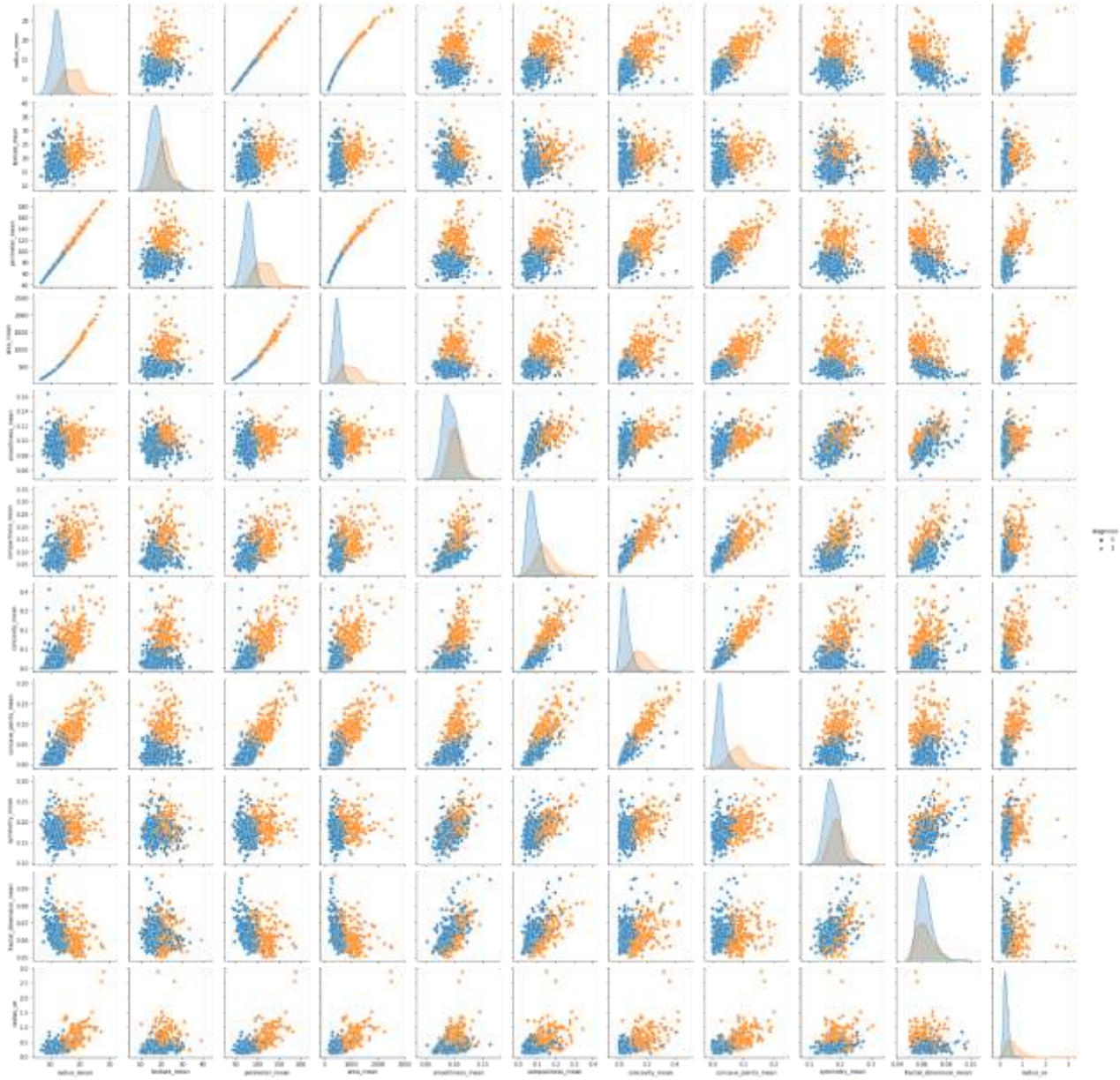
Table 1. Classification Report on Train

a. TRAIN	precision	recall	f1-score	support
0	1.00	1.00	1.00	252
1	1.00	1.00	1.00	146
accuracy			1.00	398
macro avg	1.00	1.00	1.00	398
weighted avg	1.00	1.00	1.00	398

Table 2. Classification Report on Test

b. TEST	precision	recall	f1-score	support
0	0.90	0.94	0.92	105
1	0.90	0.83	0.87	66
accuracy			0.90	171
macro avg	0.90	0.89	0.89	171
weighted avg	0.90	0.90	0.90	171

Fig.4 Decorrelation of different trees



2.4. Ensemble Methods: In an AI technique called ensemble studying, many models are coordinated to solve a comparable problem and achieve better outcomes.The key supposition is that we can identify more unambiguous areas of strength for as well as frail models when they are firmly adopted.

2.5. Bagging:

The fitted model is similarly impacted by variability: if a different dataset had been taken into consideration, we probably would have obtained a different model. When developing a model, regardless of whether we are anticipating dealing with a request or a backslide problem, we get a capability that accepts data, yields a stock up final product, and is depicted concerning the structure dataset. Therefore, bagging is crucial: in order to create a model with a smaller change, we want to suit a few free models and "typicalise" their expectations. The problem is that the

completely independent model cannot be fitted because doing so would take a lot of data. Additionally, in order to fit independent models, we rely on the "vague features" of bootstrap tests that are provided.

2.6. KNN:

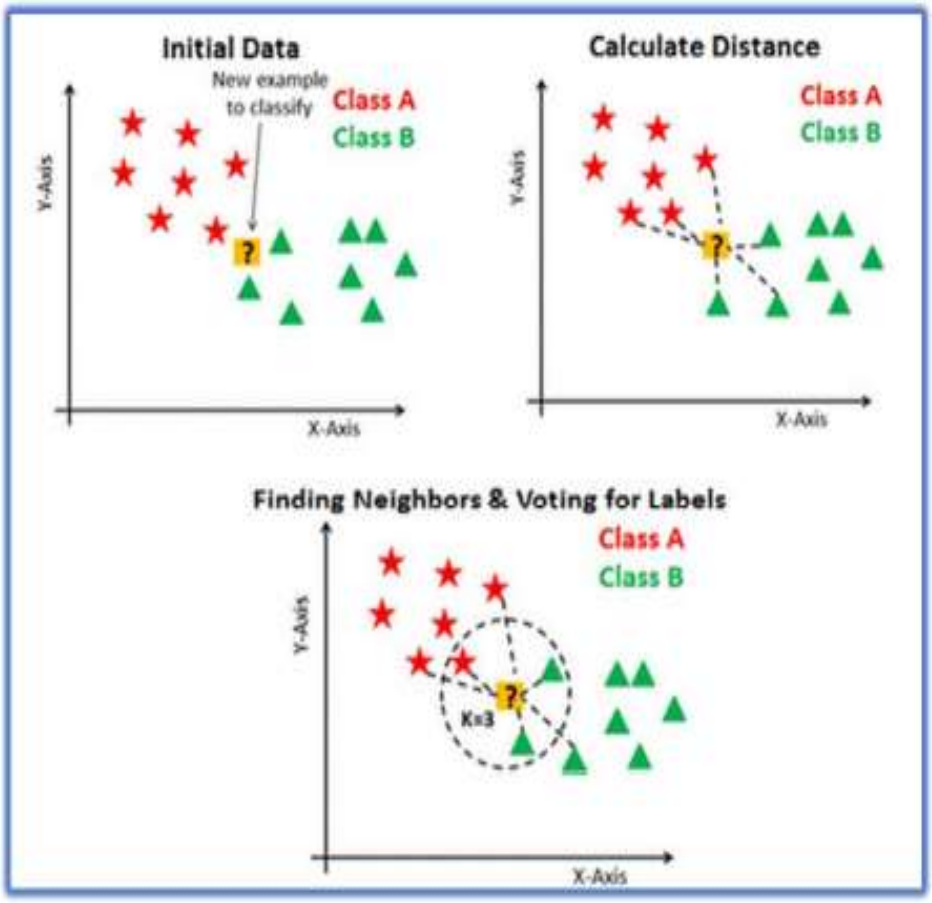
KNN, also known as K-Nearest Neighbor, is a type of AI calculation. This can be used for Characterization or Regression, but it is mainly used for grouping, so to speak. This calculation compares the new case to the group that is most fundamentally comparable to the open arrangements based on how closely the new records and available cases are related. The K-Nearest Neighbor algorithm favours both return and portrayal jobs. Characterization can be thought of as the process that organises the articles when preparing models in the component space. It lessens the significance of developing a model, altering multiple cutoff points, or similarly harbouring doubts. Euclidean distance is a mathematical condition that takes into account proximity while calculating the distance between two points on a plane.

2.6.1 Background

Assuming that A(x0, y0) and B(x1, y1) are the focus points in a plane, the Euclidean distance may then be determined at that location as-

$$\sqrt{}[(\boldsymbol{Xo}-\boldsymbol{X1})^2+\quad(\boldsymbol{Yo}-\boldsymbol{Y1})\ ^2\ ]$$

Fig.5 K-nearest neighbor



A distance metric is used to determine which of the K examples in the organisation dataset resembles another set of data. The most often used distance measurement for approved, respected input factors is Euclidean distance.

Fig.6 Confusion Matrix of KNN

Following advances should be followed during the K-NN computation

- Disparate planning data and test data.
- Cast some value of K.
- Determine a specific distance measurement.
- Embrace a model on requested test data, then record the distance of the model's n planning tests.
- Categorize determined distance from K-nearest test data.
- Grant test class to those who receives majority votes.

Significant tuning boundaries for KNN can given as follow:

n-neighbours - gravitate the quantity of closest neighbours i.e.K in the KNN computation.

Loads - employs this potential in expectation.

Fig.7 Code Snippet

```
param_grid = {'n_neighbors':[3,4,5,6,7,8
    ,9,10,11,12], 'weights': ['uniform',
    'distance']}
knn = GridSearchCV(KNeighborsClassifier
    (), param_grid = param_grid, cv=5,
    scoring = 'f1_weighted')
knn.fit(std_data_train, train_y)
Final accuracy Score Train Data: 0
    .984924623115579
Test Data: 0.9239766081871345
```

3. PROPOSED WORK

3.1 Decision Tree

One of the most common controlled learning techniques is the use of decision trees. In directed learning, your continuous data is checked and you are already aware of which lead you want to anticipate in the new information you receive, as opposed to solo realising (where data is analysed using estimations to find plans and there is no closure outcome variable to coordinate the creative process)[2][5]. This type of calculations is used by autonomous vehicles to recognise pedestrians and objects, or by businesses to gauge the lifetime value and retention rates of their clients.

Decision Trees are AI algorithms that consistently divide informational collections into more noticeable undetectable data clusters based on an obvious component, until they reach units that are sufficiently small to be handled through some imprint. They try to label new data with the aid of your named data, which they assume you already have (put apart with at least one name, such as the creature name in the images of creatures)[1][5]. These calculations are fantastic for addressing issues with affiliation (where machines classify information, eventually determining if an email is malicious) and backslide (where machines separate values, akin to a property cost)[2][3].

Characterization trees are used when the contingent variable is conventional or quantitative, while backslide trees are utilised when the limited variable is evident or emotive (such as supposing we have any desire to assess the risk that a customer will default on a development) (for instance to explore the blood grouping of a person)[1][4]. Decision trees' importance stems from the fact that they serve a variety of functions. They are applied to numerous capabilities in certain pursuits and are arguably the most intricate AI estimations.

3.1.1 BACKGROUND

Important tuning parameters for Decision trees are given as following-

- Min Sample Leaf -minimum stripped samples that is required to built a leaf node in order to smoother the model.
- Max Depth-height of tree.
- Min sample split-least samples required to split internal node
- Max leaf nodes-no.of features for best split
- criterion-measure of quality of split.

Fig.8 Code snippet of decision tree

```
11  param_grid = {'max_depth': np.arange(3, 5
        ), 'max_features': np.arange(3,5)}
12
13  tree = GridSearchCV
        (DecisionTreeClassifier(), param_grid
        , cv = 5)
14
15  tree.fit( train_x, train_y )
16
17  The overall accuracy score for the Train
        Data is: 0.9572864321608O4
18
19  The overall accuracy score for the Test
        Data is: 0.8947368421052632
20
```

4.RESULT AND DISCUSSION

It has been seen that the smoothness and working performance of the model that is generated by the combination of major machine learning techniques providing more smooth and faster result than other .

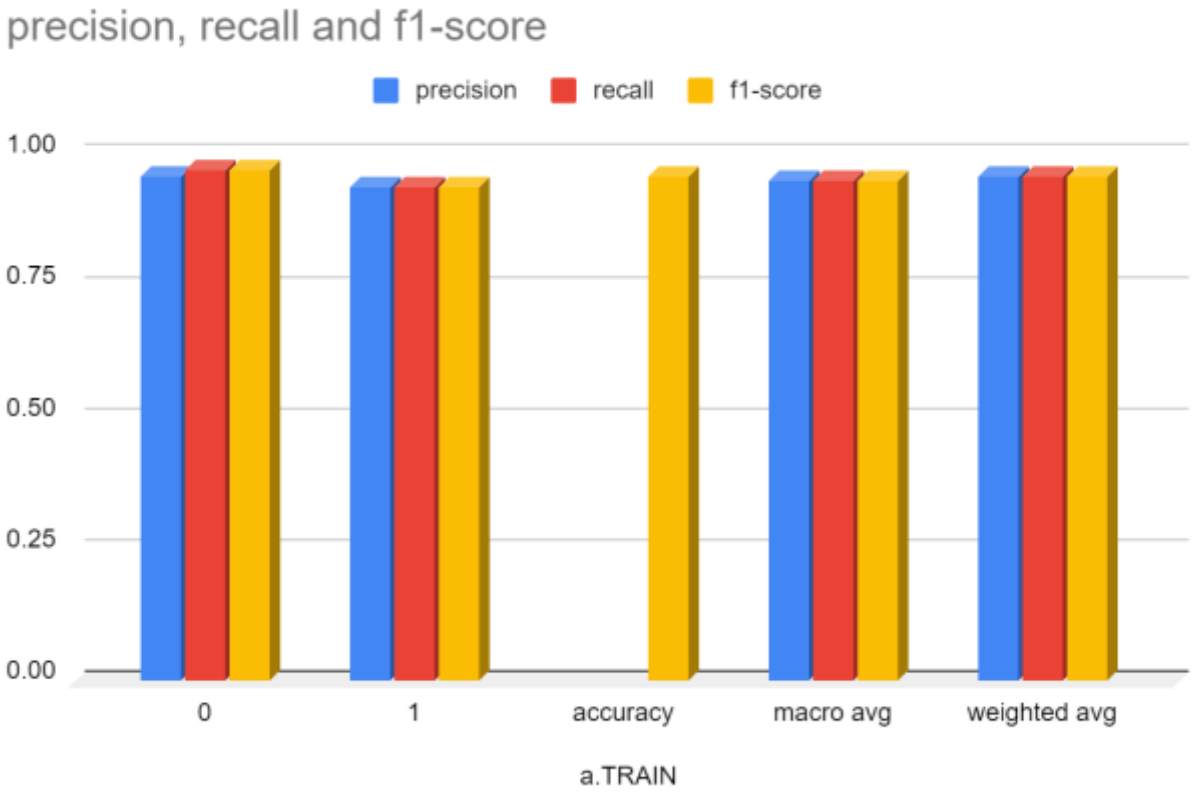
4.1 Performance Analysis

The performance of the model by the combination of the Decision tree, Artificial Neural network K-nearest neighbor & Support Vector Machine performing more reliably and smoothly than the model proposed by other Machine Learning techniques.Table 3 on comparison to Table 4 gives more accurate and reliable result as far as concerned.

TABLE 3 Classification report on train data

a.TRAIN	precision	recall	f1-score	support
0	0.96	0.97	0.97	252
1	0.94	0.94	0.94	146
accuracy			0.96	398
macro avg	0.95	0.95	0.95	398
weighted avg	0.96	0.96	0.96	398

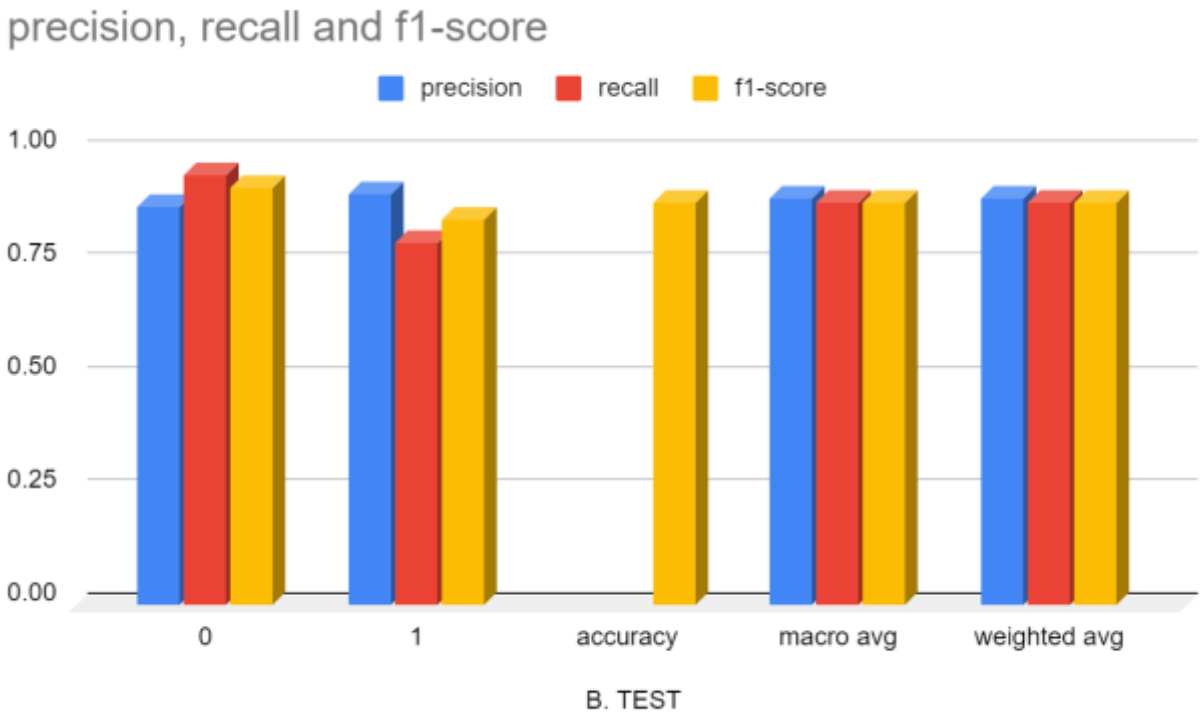
Below we are depicting a bar graph on training data to insure that the proposed model must attain a maximum amount of certainity to one another and gives out the best possible results.



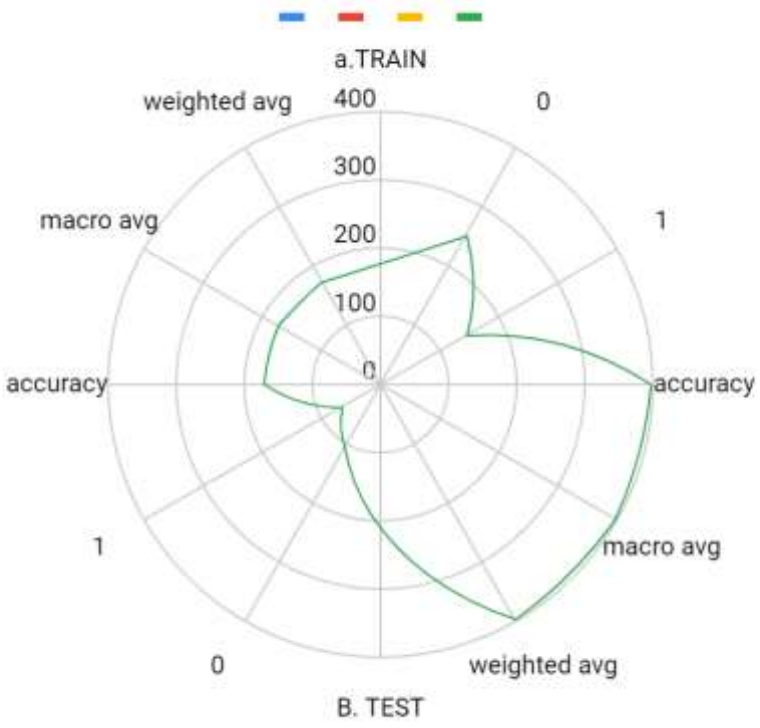
Classification report developed on the basis of test data in order to accumulate the best behavior of the proposed model that is generated after the combination of different machine learning techniques such as knn, scm, decision tree etc.

TABLE 4 Classification report on test data

B. TEST	precision	recall	f1-score	support
0	0.88	0.95	0.92	105
1	0.91	0.80	0.85	66
accuracy			0.89	171
macro avg	0.90	0.89	0.89	171
weighted avg	0.90	0.89	0.89	171



On Making a comparitive analysis on both classification report of training and test data we came to a conclusion that the model that is proposed provide more accurate and close result and also the efficiency of the model is high.



Hence, the detection of malignant tissue in patients using machine learning techniques are providing more accurate and reliable results more economically and technically. and most of the patients are getting treatment on time as they are able to detect the disease as soon as it's known.

5. Conclusion

This undertaking is mostly founded on the progression of prescient models to accomplish extraordinary accuracy in anticipating the authentic illness results utilizing regulated AI calculations. The examination of the outcomes reasons that the juxtaposition of complex information with different characterization, highlight determination, and dimensionality decrease strategies can give valuable hardware to derivation in this area. For the better use of the collection approaches and the ability to anticipate more factors, much more exploration in this area needs to be completed.This dataset includes 32 distinct components that help to condense the enormous, multifaceted dataset into a few key features. The K-Closest Neighbor (KNN), Backing Vector Machine (SVM), and Strategic Relapse, out of all the calculations used, contribute the most notable precision of 92.7% when viewed differently in relation to other estimations. As a result, in keeping with this, we may suggest that SVM is the best strategy for selecting the assumption for a Bosom Malignant growth event with complex datasets.

5.1 Authors Contributions

Conceived and analysis design were carried out by VV AND VJ. While, the data is figured and collected in contribution by VR and VW. VJ drafted the article and carefully analysed the intellectual data.VJ and VV work on working and testing of the proposed model.

6. References

1. UK Statistics Authority. Cancer statistics registrations: registrations of cancer diagnosed in 2004, England. London: UK Statistics Authority; 2007. (Series MB1 No. 35).
2. Birmingham Research Unit. Weekly returns survey- Annual Prevalence Report 2007. Royal College of General Practitioners;
3. Musa M. Treatment and outcomes for high-risk and metastatic breast cancer in California: an inquiry into disparities and research needs. California Breast Cancer Research Program. 2004
4. Garvican L, Littlejohns P. Comparison of prognostic and socio-economic factors in screen-detected and symptomatic cases of breast cancer. Public Health. 1998;112(1):15–20.
5. P. Boix-Montesinos, M.J. Vicent,, A. Armiñán, M. Orzáez, P.M. Soriano-Teruel. The past, the present, and the future of breast cancer models for nanomedicine development Adv. Drug Deliv. Rev., 173 (2021), pp. 306-330
6. Boost the Performance of Deep Learning Model Based on Real-Time Medical Images." Journal of Sensors 2023 (2023).
7. Babu, S. Z., et al. "Abridgement of Business Data Drilling with the Natural Selection and Recasting Breakthrough: Drill Data With GA." Authors Profile Tarun Danti Dey is doing Bachelor in LAW from Chittagong Independent University, Bangladesh. Her research discipline is business intelligence, LAW, and Computational thinking. She has done 3 (2020).

8. NARAYAN, VIPUL, A. K. Daniel, and Pooja Chaturvedi. "FGWOA: An Efficient Heuristic for Cluster Head Selection in WSN using Fuzzy based Grey Wolf Optimization Algorithm." (2022).
9. Faiz, Mohammad, et al. "IMPROVED HOMOMORPHIC ENCRYPTION FOR SECURITY IN CLOUD USING PARTICLE SWARM OPTIMIZATION." Journal of Pharmaceutical Negative Results (2022): 4761-4771.
10. Narayan, Vipul, A. K. Daniel, and Pooja Chaturvedi. "E-FEERP: Enhanced Fuzzy based Energy Efficient Routing Protocol for Wireless Sensor Network." Wireless Personal Communications (2023): 1-28.
11. Tyagi, Lalit Kumar, et al. "Energy Efficient Routing Protocol Using Next Cluster Head Selection Process In Two-Level Hierarchy For Wireless Sensor Network." Journal of Pharmaceutical Negative Results (2023): 665-676.
12. Paricherla, Mutyalaiah, et al. "Towards Development of Machine Learning Framework for Enhancing Security in Internet of Things." Security and Communication Networks 2022 (2022).
13. Sawhney, Rahul, et al. "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease." Decision Analytics Journal 6 (2023): 100169.
14. Srivastava, Swapnita, et al. "An Ensemble Learning Approach For Chronic Kidney Disease Classification." Journal of Pharmaceutical Negative Results (2022): 2401-2409.
15. Mall, Pawan Kumar, et al. "FuzzyNet-Based Modelling Smart Traffic System in Smart Cities Using Deep Learning Models." Handbook of Research on Data-Driven Mathematical Modeling in Smart Cities. IGI Global, 2023. 76-95.
16. Mall, Pawan Kumar, et al. "Early Warning Signs Of Parkinson's Disease Prediction Using Machine Learning Technique." Journal of Pharmaceutical Negative Results (2022): 4784-4792.
17. Pramanik, Sabyasachi, et al. "A novel approach using steganography and cryptography in business intelligence." Integration Challenges for Analytics, Business Intelligence, and Data Mining. IGI Global, 2021. 192-217.
18. Narayan, Vipul, et al. "Deep Learning Approaches for Human Gait Recognition: A Review." 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE, 2023.
19. Narayan, Vipul, et al. "FuzzyNet: Medical Image Classification based on GLCM Texture Feature." 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE, 2023
20. Mahadani, Asim Kumar, et al. "Indel-K2P: a modified Kimura 2 Parameters (K2P) model to incorporate insertion and deletion (Indel) information in phylogenetic analysis." Cyber-Physical Systems 8.1 (2022): 32-44.
21. Singh, Mahesh Kumar, et al. "Classification and Comparison of Web Recommendation Systems used in Online Business." 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM). IEEE, 2020.
22. Awasthi, Shashank, Naresh Kumar, and Pramod Kumar Srivastava. "A study of epidemic approach for worm propagation in wireless sensor network." Intelligent Computing in Engineering: Select Proceedings of RICE 2019. Springer Singapore, 2020.

23. Srivastava, Arun Pratap, et al. "Stability analysis of SIDR model for worm propagation in wireless sensor network." Indian J. Sci. Technol 9.31 (2016): 1-5.
24. Ojha, Rudra Pratap, et al. "Global stability of dynamic model for worm propagation in wireless sensor network." Proceeding of International Conference on Intelligent Communication, Control and Devices: ICICCD 2016. Springer Singapore, 2017.
25. Shashank, Awasthi, et al. "Stability analysis of SITR model and non linear dynamics in wireless sensor network." Indian Journal of Science and Technology 9.28 (2016)