# IMPROVEMENT IN AUTOMATED DIAGNOSIS OF LIPOSARCOMA USING MACHINE LEARNING

S.Mahammad Rafi[1], K.Bhavya Sri[2], E.Hemalatha[3], D.Charitha[4], H.HarshaVardhan Raju[5], K.Anil[6]

[1]*Assistant Professor in Department of Computer Science and Engineering, Annamacharya Institute of Technology and Sciences(Autonomous), Rajampet, Andhra Pradesh, India.*

[2,3,4,5,6]*Department of Computer Science and Engineering, Annamacharya Institute of Technology and Sciences(Autonomous), Rajampet, Andhra Pradesh, India.*

## ABSTRACT

*Sarcomas in the form of soft tissue tumours (STT) can develop in the connective, encircling, and supporting tissues of the body. When seen by Magnetic Resonance Imaging, they appear to be heterogeneous due to their shallow frequency in the body and their great diversity. They are frequently confused with other illnesses such lymphadenopathy, struma nodosa, and fibro adenoma mammae, and these diagnostic mistakes have a significant negative impact on the medical treatment of patients. Numerous machine learning models have been put forth by researchers to categorise cancers, but none have sufficiently addressed the issue of incorrect diagnoses. Additionally, comparable studies that have suggested methods for evaluating these tumours typically do not take the heterogeneity and magnitude of the data into account. For this reason, we suggest a machine learning-based strategy that combines a novel method of preprocessing the data for feature transformation, resampling methods to remove bias and the deviation of instability, and performing classifier tests using the Support Vector Machine (SVM) and Logistic Regression algorithms (LR). Tests conducted on data gathered at Yogyakarta, Indonesia's Nur Hidayah Hospital, reveal a significant advancement over earlier research. These findings support the idea that machine learning techniques could offer practical and useful tools to support STT diagnostics' automatic decision-making procedures.*

**Keyword:** *Support Vector Machine, LogisticRegression, Random Forest.*

## I. INTRODUCTION:

The word "soft tissue" describes tissues such as fat, muscles, and blood vessels, deep cutaneous tissues, nerves, and tissues lining joints that support, connect to, or surround other bodily structures and organs. These delicate tissues, as their name implies, are susceptible to a variety of diseases, including tumours, which can grow practically anywhere on the human body. Because they have a lot in common on a microscopic level, show comparable symptoms, and respond to treatment practically identically, the malignant varieties of these tumours, also known as Soft Tissue Sarcomas (STS), are categorised together. However, due to the difficulties in finding these tumours, efficient detection of Soft Tissues Tumors (STT) is still a significant challenge. In order to improve the detection of such malignancies, a number of approaches have been developed, including MRI analysis. With well-known biological characteristics including cellular origins and tumour specimens utilised to identify tumours, MRI is currently regarded as the gold standard diagnostic method for the detection and classification of STT. For several reasons, including ease of computation, extensive correlation between textural characteristics and tumour pathology, and robustness to changes in MRI acquisition parameters such as changes in the resolution of the tumour

image and the corruption of the MRI image caused by heterogeneity of the magnetic field, MRI can be used to analyse textural characteristics or other less well-characterized tumour characteristics. MRI's high magnetic field heterogeneity makes it challenging to distinguish the texture of some malignant tumours.
Additionally, individuals have a limited ability to distinguish between different textures. As a result, machine learning (ML) algorithms are being used increasingly frequently to interpret MRI pictures more effectively and to automatically identify tumours. Predictive automatic learning algorithms have reinforced it and made it a more crucial tool for modern medicine. These algorithms help existing expert systems make diagnoses more effectively. We have created a machine learning-based method for the automobile detection and diagnosis of tumours like STT among these several applications. Malignant tumours called STT can grow inside of blood vessels, muscles, fat, nerves, and fibrous structures. These difficulties have halted the development of new medicinal drugs due to their low frequency and the difficulty doctors have evaluating outcomes. Additionally, it is challenging for doctors to choose an appropriate course of treatment because of the variable MRI findings. STT can also be mistaken for conditions like lymphadenopathy, struma nodosa, and fibro adenoma mammae. The patient's treatment is significantly impacted by this diagnostic failure. According to Karanian and Coindre's view, benign lesions, tumours with local potential, tumours with minimal metastatic potential, and sarcomas are the four categories of connective tumour progression.

Once a molecular anomaly of an entity has been discovered, the entity's histology and molecular definitions can be found. Therefore, the present problem is to effectively employ the characteristicsthe predictive identification of STT is enhanced by the application of classification approaches and required to prevent delays in identifying the patient and optimising their treatment. Because of this, Nur Hidayah Hospital estimates that less than 1% of adult cancers—which have a lifetime chance of development of 0.33%—develop in soft tissues such fat, muscle, nerves, fibrous tissue, and blood vessels. As a result, STT have had a sense of mystery surrounding their diagnosis of being thought uncontrollable for decades. They were initially distinguished by their typically poor prognosis and restricted treatment options. Indeed, the use of cytogenetics and molecular biology in the diagnosis of STT has resulted in significant improvements in investigative techniques. Additionally, the conventional histological frameworks have been reexamined in light of the identification of recurrent aberrations in other fields of disease. A new categorization of soft tissue tumours by the World Health Organization that takes into consideration genetic and molecular information was born in 2002 as a result of these advancements. According to the following categories, STT types were categorised in this edition: adipocytic, fibroblastic/myofibroblastic, fibrohistiocytic, smooth muscle, pericytic/perivascular, skeletal muscle, vascular, chondro-osseous tumours, and the category known as "uncertain differentiation tumours." The fourth version of the WHO's categorization of STT, which was created based on the type of tumour as well as the morphological, immunohistochemical, and genetic characteristics, was published in February 2013—eleven years after the third edition. The 2013 edition also contains chapters on nerve sheath tumours and gastrointestinal stromal tumours, as well as a newly added section on undifferentiated/unclassified sarcomas.

To provide patients with the best care, the WHO categorization has made it possible to diagnose particular cancer types more accurately. Artificial intelligence is now being used more and more in this diagnosis to diagnose malignancies more accurately using ML algorithms. The difficulties of applying ML to the categorization of STT will be covered in the section that follows. In order to deliver effective therapy, Yogyakarta, Indonesia has been interested in predicting whether a patient is accurately diagnosed with the STT or non-STT. To do this, we examine a dataset of 50 patients with the STT and 25 patients who received the STT diagnosis in error. All patients had successfully completed blood coagulation and complete blood counts, and their total protein and alburnin/globumin antigen tests came back negative. These additional criteria are also included in the dataset. Other individuals who did not have the STT but who could have been misdiagnosed with it had other conditions such fibroadenoma mammae, lymphadenopathy, and struma nodosa.

Our study's primary goal is to construct and evaluate machine learning-based models with the necessary tools to allow users to extensively assess patient data and accurately distinguish STT from non-STT. The method we developed aims to address the issues with the predicted detection of STT that were present in earlier work. The remainder of the essay is structured as follows: In Section 2, we outline the characteristics of STT and the difficulties in using ML to classify STT. In Section 3, we outline a methodical and thorough procedure for developing and testing an automatic learning classifier based on the Support Vector Machine (SVM) and Logistic Regression algorithms that can be applied in real-world scenarios with almost the same performance. The complexity of classifiers, the impact of learning dataset size on classifier behaviour, and the ideal size of the training data that can be used to train a model of classifier and obtain excellent generalisation performance on invisible data are some additional questions that our research in this section will address. The data we have obtained are evaluated and analysed in Section 4, and our job is concluded in Section 6.

## II. LITERATURE SURVEY:

**[1] Annals of Translational Medicine, vol. 22, no. 3, p. 368, 2015. T. Hayashi, A. Horiuchi, K. Sano, Y. Kanai, N. Yaegashi, H. Aburatani, and I. Konishi, "Biological Characterization of Soft Tissue Sarcomas."**
Neoplastic cancers known as soft tissue sarcomas generally develop in tissues with mesenchymal origin. Numerous important aspects have made it difficult to identify novel molecular pathways causing mesenchymal transition, build new therapeutics, and develop diagnostic biomarkers. First off, less than 15,000 new instances of malignant soft tissue sarcomas are identified in the United States each year. Soft tissue sarcomas are exceedingly heterogeneous tumours that develop in a variety of tissues from a wide range of cell lineages, which further complicates matters. Clinical materials are scarce, and their intrinsic variety makes for a difficult experimental setting for researchers and physicians. In comparison to other malignant tumours, there has been very little development in the clinical therapy choices available to patients due to these difficulties. Scientists are currently using mice models whose genomes have been specially engineered to carry gene deletions, gene amplifications, and somatic mutations typically reported in human soft tissue sarcomas to gain insight into the pathobiology of soft tissue sarcomas. Through the use of these model species, we have been able to learn more about how changes in the interferon (IFN), tumour protein 53 (TP53), and/or retinoblastoma (RB) signalling pathways directly affect sarcomagenesis. Many people in the physiological community hope that using a variety of mice models could help us better understand sarcomagenesis and possibly find novel diagnostic biomarkers and treatment approaches for human soft tissue sarcomas.

**[2] B. Julesz, E. N. Gilbert, L. A. Shepp, and H. L. Frisch, Inability of humans to discriminate between visual textures that agree in second-order statistics—Revisited, Perception, vol. 2, no. 4, pp. 391–405, 1973.**

When generating pairs of random textures side by side in a previous research by Julesz (1962), did you use a Markov process with distinct third-order joint-probability distributions but the same first- and second-order distributions? The human visual system could not distinguish between such texture pairs without close inspection. The general processes underpinning visual texture discrimination are two-dimensional, whereas Markov systems are regrettably fundamentally one-dimensional. This article introduces three novel techniques for creating two-dimensional non-Markovian textures with distinct third-order statistics but the same first- and second-order statistics. All three techniques produce texture pairs that are indistinguishable from one another.

**[3]H. Farhidzadeh, B. Chaudhury, M. Zhou, D. B. Goldgof, L. O. Hall, R.A.Gatenby, R. J. Gillies, and M. Raghavan, Prediction of treatment outcome in soft tissue sarcoma based on radiologically defined habitats, in Proc. SPIE9414, Medical Imaging 2015: Computer-Aided Diagnosis, Orlando, FL, USA, 2015, p. 94141U.**

A rare and extremely aggressive variety of soft tissue sarcomas is adult-type fibrosarcoma. Its diagnosis is always one of exclusion because other sarcomas with spindle-cell shapes exist. Similar tumour entities have a significant risk of being misdiagnosed, which frequently results in ineffective tumour treatment. Here, we list the salient characteristics of fibrosarcoma. The patient's prognosis is typically quite bad when fibrosarcoma is correctly diagnosed. Low susceptibility to radiotherapy and chemotherapy as well as a high risk of tumour recurrences are characteristics of fibrosarcoma. Therefore, it is crucial to find innovative approaches to enhance the therapy of this tumour entity. We highlight several exciting new lines of inquiry into fibrosarcoma, concentrating on more potent ways to target the tumour microenvironment. Cancer development, invasion, metastasis, and chemosensitivity all depend heavily on communication between tumour cells and the stromal tissue that surrounds them. Targeting the tumour microenvironment has therapeutic promise, which is discussed.

**[4] G. Marinakos and S. Daskalaki, Imbalanced customer classification for bank direct marketing, J. Mark. Anal., vol. 5, no. 1, pp. 14–30, 2017.**

This study suggests a data pre-processing strategy for a neural network for binary classification used in bank telemarketing. 19 features, 16 features, and 20 features from three datasets have been used to assess how well the algorithm performs in relation to the categorization model. The three stages of the data pre-processing procedure are data cleaning, data imbalance correction, and lastly data normalisation. The results of this study showed that a binary classification model combined with data cleaning methods like Missing Common (MC) and Tomsk Links

(TL) performs better than Ignore Missing (IM). MaxAbsScaler (MAS) and MinMaxScaler (MMS) have consistently shown to perform better than other normalising algorithms in terms of data normalisation. The classification model used in this paper makes use of the MC-TL-MMS data pre-processing method combination. By using 16 features and 20 features, respectively, the algorithm using this method was able to record an area of the receiver operating characteristic curve (AUC) of 0.9129 and 0.9464. Compared to other earlier studies, this result shows the highest performance accuracy figure.

**[5] S. L. Salzberg, Book review: C4.5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann publishers, inc., 1993, Mach. Learn, vol. 16, no. 3, pp. 235–240, 1994.**

In addition to discussing some suggestions for enhancing the capabilities of the algorithm, Quinlan examines some of C4.5's drawbacks, such as its bias towards rectangular regions. He does not examine competing decision tree construction algorithms or competing classification methods, with the exception of a brief overview in the first chapter, likely because the book is not meant to serve as a general study of learning algorithms. The book should be most helpful as a research tool for machine learning experts or as a supplemental material in a graduate or advanced undergraduate course. Quinlan does not presume that readers are familiar with any prior work on decision trees and thoroughly explains the necessary principles in the first few chapters.

**[6] F. Collin, M. Gelly-Marty, M. B. N. Binh, and J. M. Coindre, Sarcomes des tissus mous: Donne´es anatomopathologiquesactuelles, Cancer/Radioth´erapie, vol. 10, nos. 1&2, pp. 7–14, 2006.**

Pathology has changed significantly in the area of soft tissue tumours over the past fifteen years. They were connected to important developments in genetics and molecular biology. The WHO classification was updated as a result of new information. The term "malignant fibrous histiocytoma" is no longer used. It has divided into many subtypes that are either undifferentiated sarcomas, liposarcomas, or leiomyosarcomas. Reevaluation led to the classification of haemangiopericytoma as a solitary fibrous tumour. Many tools have been enhanced. With the development of new antibodies, immunohistochemistry's specificity increased and in some circumstances, such as the detection of c-kit proto-oncogen mutations in gastrointestinal stromal tumours by CD117, it became useful for predicting therapeutic response. From core needle biopsies, improved techniques enable precise diagnoses. Surgeons and pathologists jointly analyse surgical specimens, paying close attention to the resection margins. The French Federation of Cancer Centers' classification system continues to be the strongest predictor of metastasis-free survival and overall patient survival, despite some restrictions. It is based on an evaluation of three factors: differentiation, the degree of necrosis, and the number of tumours that have undergone mitosis. The pathologist establishes a diagnosis and actively participates in the prognosis and therapeutic response prediction. He plays a significant role in the decision-making process for the multimodal treatment of sarcomas.

## III. ALGORITHMS:

### 1. Logistic Regression:

One of the most often used Machine Learning algorithms, within the category of Supervised Learning, is logistic regression. Using a predetermined set of independent factors, it is used to predict the categorical dependent variable. In a categorical dependent variable, the output is predicted via logistic regression. As a result, the result must be a discrete or categorical value. Rather of providing the exact values of 0 and 1, it provides the probabilistic values that fall between 0 and 1. It can be either Yes or No, 0 or 1, true or false, etc. With the exception of how they are applied, logistic regression and linear regression are very similar. While logistic regression is used to solve classification difficulties, linear regression is used to solve regression problems.

In logistic regression, we fit a "S" shaped logistic function, which predicts two maximum values, rather than a regression line. The logistic function's curve shows the possibility of several things, including whether or not the cells are malignant, whether or not a mouse is obese depending on its weight, etc. Because it can classify new data using both continuous and discrete datasets, logistic regression is a key machine learning approach. When classifying observations using various sources of data, logistic regression can be used to quickly identify the factors that will work well.

The logistic function is displayed in the graphic below. A dataset is provided that includes data on various users collected from social networking sites. A new SUV vehicle was just introduced by an automobile manufacturer. The business therefore needed to determine how many consumers in the dataset were interested in buying a car. Using the logistic regression approach, we will create a machine learning model for this issue. The graphic below displays the dataset. We will use age and salary to forecast the purchased variable in this problem.

**2. Support Vector Classifier:**

SVMs are capable of handling both classification and regression issues. The decision boundary for this method's hyperplane needs to be defined. A decision plane is required to divide a collection of objects into their many classes. If the objects cannot be separated linearly, kernels—complex mathematical functions—must be used to separate the objects that belong to various classes. The goal of SVM is to correctly identify the objects using examples from the training data set. These are some benefits of SVM: It can manage structured and semi-structured data, and if the right kernel function can be determined, it can manage complex functions.

Less likelihood of overfitting exists since SVM adopts generalisation. With large-scale data, it can scale up. It does not become trapped in regional optimum. The following are SVM's drawbacks: due to the longer training times required for large data sets, its performance suffers. Finding an adequate kernel function will be challenging. SVM performs poorly when the dataset is noisy. SVM doesn't offer probabilities in its output. It's challenging to comprehend the final SVM model.

Support The practical use of vector machines includes text classification, handwriting identification, face detection, credit card fraud detection, and cancer diagnosis. Therefore, the first technique to try will be the logistic regression approach, followed by the decision trees (Random Forests) to see if there is a noticeable improvement. The third approach to try is the SVM strategy. SVM can be tested when there are lots of observations and features.

**3. Random Forest:**

A random forest is a machine learning method for tackling classification and regression issues. It makes use of ensemble learning, a method for solving complicated issues by combining a number of classifiers.

In a random forest algorithm, there are many different decision trees. The random forest algorithm creates a "forest" that is trained via bagging or bootstrap aggregation. The accuracy of machine learning algorithms is increased by bagging, an ensemble meta-algorithm.

Based on the predictions of the decision trees, the (random forest) algorithm determines the result. It makes predictions by averaging or averaging out the results from different trees. The accuracy of the result grows as the number of trees increases.

The decision tree algorithm's shortcomings are eliminated with a random forest. It improves precision and decreases dataset overfitting. Without requiring numerous configurations in packages, it generates forecasts (like Scikit-learn).

## IV. EXPERIMENTAION RESULTS:



## V. CONCLUSION:

Applications in many fields, including medicine, can benefit from high precision calculations enhanced by ML algorithms. The performance of computer-aided diagnostic systems has improved greatly thanks to these technologies in recent years, but integrating them continues to be difficult for contemporary healthcare organisations. Based on data gathered from the Nur Hidayah Hospital in Bantul, Yogyakarta, Indonesia, we built a solid and realistic model in this study that enables automatic predictive classification of STT and non-STT. After including a fresh data pretreatment method, we contrasted the SVM and LR classifiers. This comparison revealed that the LR model is substantially more sensitive to the amount of variables than the SVM model, even though the LR algorithm is marginally more efficient than the SVM algorithm.

## VI. REFERENCES:

[1] F. Collin, M. Gelly-Marty, M. B. N. Binh, and J. M. Coindre, "Sarcomas of the Muscle Tissue: Current Anatomicopathological Reports," Cancer/Radiotherapy, vol. 10, nos. 1 & 2, pp. 7–14, 2006.

[2] Biological characterization of soft tissue sarcomas, Annals of Translational Medicine, vol. 22, no. 3, p. 368, 2015. T. Hayashi, A. Horiuchi, K. Sano, Y. Kanai, N. Yaegashi, H. Aburatani, and I. Konishi.

[3] S. L. Salzberg, book review of J. Ross Quinlan's C4.5: Programs for Machine Learning. Machine Learning, vol. 16, no. 3, 1993, pp. 235–240, Morgan Kaufmann Publishers, Inc.

[4] Inability of humans to differentiate between visual textures that agree in second-order statistics—Revisited, Perception, vol. 2, no. 4, pp. 391-405, 1973. B. Julesz, E. N. Gilbert, L. A. Shepp, and H. L. Frisch.

[5] Prediction of treatment outcome in soft tissue sarcoma based on radiologically defined habitats, Proc. SPIE 9414, Medical Imaging 2015: Computer-Aided Diagnosis, Orlando, FL, USA, 2015, p. 94141U. H. Farhidzadeh, B. Chaudhury, M. Zhou, D. B. Goldgof, L. O. Hall, R. A. Gatenby, R. J. Gillies.

[6] Imbalanced client classification for bank direct marketing, G. Marinakos and S. Daskalaki, J. Mark. Anal., vol. 5, no. 1, pp. 14–30, 2017.

[7] Application of improved decision tree approach based on rough set in developing smart medical analysis CRM system, International Journal of Smart Homes, vol. 10, no. 1, pp. 251-266, 2016. H. S. Xu, L. Wang, and W. L. Gan.

[8] Asymptotic behaviour of support vector machines with Gaussian kernels, Neural Comput., vol. 15, no. 7, pp. 1667–1689, 2003. S. S. Keerthi and C. J. Lin.

[9] Infinite-limits for Tikhonov regularisation, by R. A. Lippert and R. M. Rifkin, J. Machine Learning Research, vol. 7, pp. 855-876, 2006.

[10] C. G. L. Guillou, Tumeurs des tissus mous: Role of the pathologist in the diagnostic approach, Review of Medical Switzerland, vol. 3, p. 32473, 2007.