International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Real-Time Imbalance Liver Tumor Sensor Databases: A Deep Classification Framework with Ensemble Feature Extraction, Ranking, and Probabilistic Segmentation for Efficient Analysis

N. Nanda Prakash[1*], V. Rajesh[2], Sandeep Dwarkanath Pande[3], Syed Inthiyaz[4], Sk Hasane Ahammad[5], Dharmesh Dhabliya[6], Rahul Joshi[7]

Submitted: 05/02/2024 **Revised**: 13/03/2024 **Accepted**: 19/03/2024

Abstract: As the size of liver tumor image databases increases, it becomes challenging to enhance the true positive rate of traditional prediction approaches due to the high majority-minority ratio and noise. However, while 3D convolutions have the potential to fully leverage spatial information, they also come with high computational costs and require significant GPU memory usage. On the other hand, 2D convolutions are limited in their ability to utilize the information contained in the third dimension. Missing feature values, feature noise, and Imbalanced liver classes are some of the significant factors that can impact the quality of input data. The quality of imbalance data significantly impacts the efficiency of classification approaches, making it necessary to ensure high-quality input data to achieve optimal results. Therefore, to ensure high-quality predictions on imbalanced liver datasets, models need to be optimized. Sensors are commonly used to collect and measure physical parameters, and they can be used to obtain liver tumor data for the proposed model. In this work, medical imaging sensors such as CT (computed tomography) machines are used to capture detailed images of the liver and identify potential tumors. Therefore, sensors play a crucial role in the proposed model by providing the necessary data to extract features, segment the liver and detect tumors accurately. In this work, an optimized k-joint probabilistic segmentation-based ensemble classification model is proposed to address the issues of homogenous and heterogenous liver tumor detection. Additionally, novel image filtering, feature extraction and ranking approaches are proposed to improve the imbalanced liver tumor regions for classification process. The experimental results demonstrate that the proposed classification model based on k-joint probability segmentation has significantly improved the accuracy, recall, precision, and AUC compared to the existing models.

Keywords: *Imbalance liver image data, probabilistic segmentation, deep learning, support vector machine, decision tree, ensemble learning model.*

1. Introduction

Liver cancer is a severe disease that claims many lives every year. Accurate CT scans that measure the tumor's size, shape, position, and functional volume can help doctors detect and treat hepatocellular carcinoma more effectively [1]. Therefore, there is a significant requirement for automatic liver and liver tumor segmentation methods in the medical field. Automatically segmenting the liver from contrast-enhanced CT scans is challenging due to the low contrast ratio. Figure 1 displays the liver's closeness to adjacent organs. To improve tumor visibility, radiologists inject a contrast agent during CT scans, but this can create noise in the liver region [2]. Liver segmentation is already a challenging endeavor, and tumor segmentation is even more difficult. Figure 1 demonstrates how liver tumors can vary significantly in size, shape, location, and number within a single patient, posing a hurdle for automatic segmentation. Certain lesions may not possess clear boundaries, as shown in Figure 1's

third row, which can impede the effectiveness of edge-based techniques. Several methods are available for breaking down a larger entity into smaller, more manageable components, commonly referred to as segmentation techniques [3].

[1]*Department of ECE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, 522302, India.*
Email: *nandaprakashnelaturi@gmail.com*
**(Corresponding Author)*
[2]*Department of ECE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, 522302, India.*
Email: rajesh4444@kluniversity.in
[3]*School of Computer Engineering, MIT Academy of Engineering, Alandi, Pune. Dist., Pune, Maharashtra–412105, India.*
Email: sandeep7887pande@gmail.com
[4]*Department of ECE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, 522302, India.*
Email: syedinthiyaz@kluniversity.in
[5]*Department of ECE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, 522302, India.*
Email: ahammadklu@gmail.com
[6]*Professor, Department of Information Technology, Vishwakarma Institute of Information Technology, Pune, Maharashtra, India. Email: dharmesh.dhabliya@viit.ac.in*
[7]*Associate Professor, Computer Science and Engineering, Symbiosis Institute of Technology, Pune, Symbiosis International (Deemed University), Pune - 412115, India.*
Email: rahulj@sitpune.edu.in

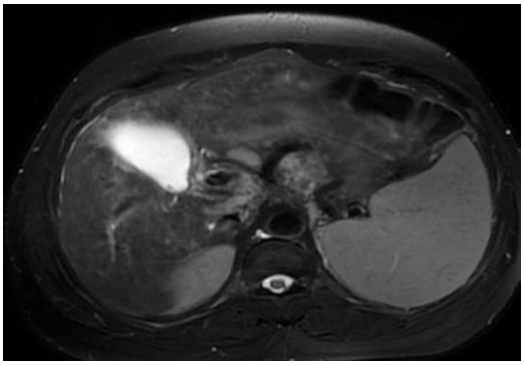


Fig. 1. Liver tumors can vary in size, shape, and location of patient

Class imbalance is a common issue in real-world applications where the distribution of classes in the dataset is uneven. For instance, medical diagnosis data may have significantly fewer instances of a specific disease than instances without the disease. Existing models may overlook the minority class, resulting in subpar performance [2], and this mismatch in data distribution can lead to biased conclusions [4]. Several methods have been proposed to address imbalanced liver datasets, including oversampling, under sampling, and cost-sensitive learning. However, these methods have limitations and often result in information loss or overfitting. They also require significant resources and are computationally expensive. Class imbalance occurs when one class has significantly more instances than the other class. Real-world applications such as credit scoring, fraud detection and medical diagnostics frequently involve this type of Imbalanced liver dataset [5]. On the other hand, class distribution imbalance describes a scenario where the number of instances of each class is not significantly different, but the distribution of the classes is unequal. Additionally, concept drift is a situation where the distribution of classes changes over time, making it difficult for machine learning models to adapt. Covariate shift occurs when the distribution of the input features varies for each class. This type of Imbalanced liver dataset arises when data is collected from multiple sources or at different times, making it challenging for machine learning models to accurately predict the class labels. Imbalanced liver datasets in machine learning algorithms can lead to poor accuracy, biased outputs, and overfitting. Existing models often overlook the minority class in imbalanced liver datasets, resulting in subpar performance and distorted outcomes. Ensemble learning, where multiple machine learning models are combined to produce a single prediction, is a useful method for improving the performance of the machine learning model. Although this approach can be computationally expensive and requires significant resources, it has shown promising outcomes in practical applications.

In addition to conventional machine learning methods, other approaches, such as deep learning and transfer learning, have been proposed to address imbalanced liver datasets. These approaches offer different strategies for handling the issue of imbalanced liver datasets in machine learning and have shown promising results in real-world applications. Various oversampling methods are available, including random oversampling, synthetic oversampling, and adaptive synthetic oversampling [6]. Resampling techniques, such as oversampling and under sampling, have been found to be effective in handling both binary and multi-class Imbalanced liver classification types. The main advantage of these resampling techniques is their independence from the underlying classifier. Empirically, pre-processing has been demonstrated to be a good way to balance the variable class distribution. Random oversampling is the technique of replicating instances of the minority class at random until the distribution of classes is balanced. Although this approach is straightforward and easy to understand, it can lead to overfitting and an increase in computation time. Synthetic oversampling involves creating synthetic instances of the minority class using techniques like bootstrapping and bagging [7]. Imbalanced liver datasets in real-time machine learning systems can result in several issues, including predictive model bias, overfitting, and metric misrepresentation. Sampling techniques, such as oversampling or under sampling, can balance the class distribution in an Imbalanced liver dataset, but they may also result in information loss and an increase in noise in the data [8]. Synthetic data creation techniques, such as SMOTE (Synthetic Minority Over-sampling Technique), can be used to produce data samples of the minority class, but they may also result in overfitting and poorer model performance. When developing and implementing machine learning models in real-time applications, it is crucial to carefully analyze these issues and take necessary action to resolve them. These are some of the most significant issues associated with imbalanced liver datasets in real-time machine learning applications. Preparing a dataset for machine learning involves several crucial processes, including data preprocessing, noise filtering, addressing missing values, and dealing with attribute noise and class noise [9].

According to the American Cancer Society (ACS), detecting and treating liver cancer early can result in improved patient outcomes. The use of shallow and deep learning methods has shown potential in identifying liver cancer at an early stage, which can lead to successful treatment. In 2016, there were 42,710 new cases of liver cancer, with 30,186 cases affecting males and 12,638 cases affecting females. Deep learning algorithms have been incorporated, resulting in significant improvements in radiography, particularly in liver segmentation. Clinicians can now make more accurate and effective decisions regarding diagnosis and treatment, leading to precise detection and early treatment of liver disease. ACS is

confident that this approach will yield better results for patients. Advanced techniques are utilized to meticulously analyze imaging data and detect concealed malignant growths in the liver. The proliferation of cancerous cells in liver tissues is the defining characteristic of hepatocellular carcinoma (HCC), which is the most widespread form of liver cancer. The World Health Organization (WHO) has reported that liver cancer was responsible for the second highest number of fatalities worldwide in 2015, with 788,000 of the 8.8 million reported deaths linked to it [10]. Maintaining a healthy liver is crucial for overall well-being and digestion. Liver cancer can be divided into primary and secondary types, with hepatocellular carcinoma and hemangiomas being common primary liver cancers. The findings of this study offer hope for improved diagnostic and treatment options for liver cancer and associated illnesses. A new computer-assisted technology can now detect liver cancer from unprocessed CT images by analyzing anomalies in the liver's textural image. Although the technique faces challenges with size, orientation, blur, and noise, it has the potential to transform the detection and treatment of liver cancer [11]. The accurate and timely detection of cancerous masses is essential for the effective treatment of liver tumors, which can be either benign or malignant. A cutting-edge tumour segmentation technology is being developed to combat the catastrophic outcomes associated with liver tumors. Early diagnosis and detection of malignant tumors is crucial to avoid negative effects. However, medical imaging presents a challenge in detecting and segmenting benign and malignant liver tumors. This study introduces a trustworthy segmentation and detection method that utilizes abdominal CT images. The process is complicated due to the noise and varying image intensity levels in CT scans of diverse patients. Nonetheless, the objective is to develop a solution that guarantees the best possible care for patients. The study aims to expand the limits of medical imaging and enhance the prognosis of individuals with liver tumors. Delineating tumors in abdominal CT scans is a difficult and complex task, further complicated by the indistinct borders of adjacent organs. The scan reveals soft tissues with similar intensity levels to the liver, like the pancreas and spleen. The human body comprises essential organs such as the liver, colon, pancreas, spleen, and abdominal wall. However, detecting liver tumors from CT scans is a difficult task due to the presence of uncertainties, noise, and variations in intensity levels between patients. To surmount these challenges, a pioneering technique has been devised that divides tumor detection into two distinct phases. This innovative methodology aims to advance the field and improve patient outcomes by carefully examining the intricacies of tumor segmentation in abdominal CT scans. The proposed method entails segmenting the liver initially and then detecting the tumor from the previously segmented liver [12]. Despite the progress made in medicine, terminal illnesses like cancer and tumors still pose a significant challenge to treatment. Nevertheless, the rapid advancement of medical knowledge provides a ray of hope, allowing for the swift and accurate identification of various ailments. However, medical errors, which can result from human weakness, can have dire consequences, even leading to death. To address this issue, advanced medical systems that incorporate AI and machine vision are being developed to reduce human error in critical situations. The medical equipment market has seen remarkable growth, mainly in the categories of diagnostic and treatment equipment. With fresh perspectives and innovative approaches, the future of medical technology looks bright [13]. The use of imaging technologies has transformed the diagnosis, treatment, and monitoring of various illnesses. Medical imaging has become an essential component of contemporary healthcare, offering a more profound insight into the intricacies of the human body. Advanced techniques like X-rays and sound reflection enable the detection of irregularities with exceptional precision and accuracy. Despite the recent progress in diagnostic software, liver cancer is still a significant threat to public health worldwide. The industry is thriving due to the high cost of sophisticated equipment [14].

Significant advancements have been made in the detection and analysis of liver function. The use of computed tomography (CT) scans has enabled medical professionals to conduct more thorough studies on liver function, with a specific focus on lesion segmentation. Lesion segmentation is a crucial aspect in the diagnosis and prognosis of liver diseases and abnormalities. The application of advanced deep learning algorithms has revolutionized radiography, bringing about a new era of precision and accuracy in liver analysis. This has created opportunities for future research in this field. A pioneering method has been developed that utilizes 3D CT scans and convolutional neural networks (CNNs) to accurately differentiate the liver from surrounding organs. This method has resulted in significant improvements in lesion segmentation, demonstrating the impressive progress in the detection and analysis of liver function [15]. The potential impact of deep learning algorithms in radiology is vast, as demonstrated by the innovative approach outlined in this research article. Using a Convolutional Neural Network (CNN), accurate classification of each slice of a 3D scan can be achieved, allowing for the removal of non-abdominal slices and isolation of the liver for further examination. This technique involves the utilization of a second CNN to separate the liver from abdominal slices, with the segmented liver slices then meticulously combined into a volume for morphological operations during post-processing. The outcomes of this method were groundbreaking, highlighting a significant advancement in liver segmentation [16-18]. The segmentation of the liver into two distinct components allows for a precise and

dependable analysis of CT scans. It is essential to conduct thorough image preprocessing to ensure that the segmentation process yields the best possible outcomes. The initial stage of liver segmentation entails a complex methodology that employs adaptive thresholding, global thresholding, and mathematical correction approaches. Morphology techniques are used to separate the liver from the surrounding abdominal organs with accuracy, making the image more straightforward. Advanced techniques like region expanding, adaptive thresholding, and fuzzy c-mean clustering are utilized to achieve this level of accuracy [19-21]. The integration of computer-aided diagnosis (CAD) is essential for achieving accurate medical imaging. Nevertheless, the existence of salt and pepper noise in CT images can considerably impair the outcomes, making pre-processing a critical phase of the process. To address this problem, the CT image is converted to grayscale. Moreover, a 3x3 median filter is utilized during pre-processing to reduce noise. The effectiveness of a cutting-edge method in achieving precise liver segmentation, a critical aspect of CAD for medical imaging, has been proven. Precise liver segmentation can result in early detection and diagnosis of tumors, which can lead to improved patient outcomes. To accomplish image segmentation, important image characteristics were considered [22].

2. Proposed Methodology

In liver image segmentation and ensemble classification, class membership identification using segmentation is a popular technique for identifying patterns and groupings in datasets with Imbalanced liver class distributions. Segmentation can assist in identifying trends within the minority class, which may not be visible when considering the entire dataset. One approach to utilizing segmentation for imbalanced liver datasets is to use segmentation algorithms to identify segments of data within the minority class in the context of liver image segmentation. These segments can be used to generate additional data points to balance the class distribution and improve the accuracy of liver image segmentation. In addition, segmentation can be used to assign class membership to the data points. This involves determining the segment to which each data point belongs and assigning it to the minority class if it falls within a segment that contains a higher percentage of minority class data points. Probabilistic models, such as Gaussian distributions, can be used to assign class membership by modelling the distribution of each class. This approach not only allows for the prediction of the class label of an observation, but also provides a probability score that represents the confidence of the prediction. To improve the statistical classification metrics of liver image segmentation, an ensemble learning framework can be implemented on the segmented data. The proposed model's overall framework filters imbalanced liver datasets using the noise filtering approach and feature ranking measure, employs the K-density probabilistic segmentation approach to determine the class membership of the filtered data, and finally implements an ensemble learning framework on the segmented data. Figure 1 illustrates the proposed model's overall framework for liver image segmentation and ensemble classification.

2.1 Liver Image Sparse Filtering

Liver image filtering can be performed to remove sparse noise from a dataset of liver images with numerical features. Sparse filtering is a useful technique for tasks such as feature selection, denoising, and compression. However, non-linear Gaussian estimation (NGE) can be used instead to estimate the parameters of a non-linear Gaussian model. This approach offers greater flexibility in expressing the relationships and underlying structure of the data compared to traditional sparse filtering, which is a linear feature selection method. By combining NGE and sparse filtering, a non-linear Gaussian model can be learned that features sparsity. This approach can improve the accuracy and relevance of liver image analysis by removing sparse noise from the dataset. The specific implementation of this process may vary depending on the dataset and the desired outcome, but the use of NGE and sparse filtering can provide an effective method for liver image filtering as shown in Figure 2.

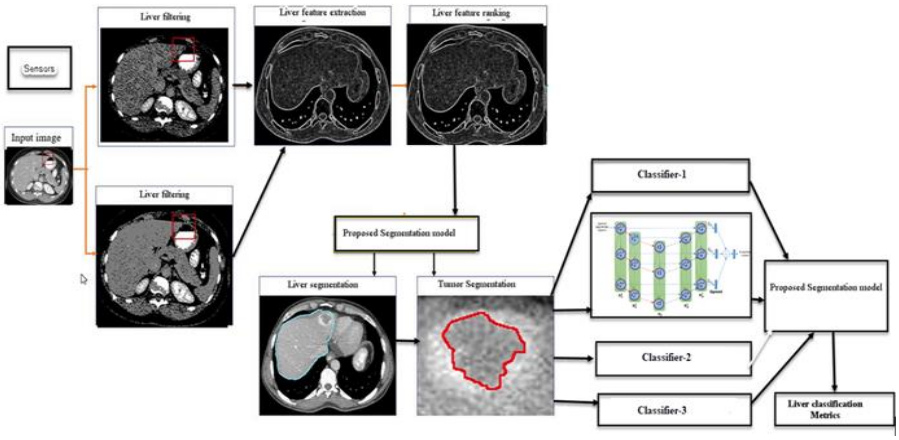


Fig. 2. Proposed Framework

To identify the sparse features in partitioned data with a high degree of missing values, which makes accurate predictions even with limited data, the proposed equation called Non-linear Gaussian estimation can be used.

2.2 Feature Extraction Measures

Feature ranking is particularly important for imbalance datasets because these datasets are characterized by a significant imbalance between the majority and minority classes, with the minority class often having a much smaller representation. Feature ranking is an essential step in the process of developing effective machine learning models for imbalance datasets. To enhance the performance of models and prevent overfitting, it is crucial to choose the most relevant features for imbalance datasets when the majority and minority classes are strongly Imbalanced liver. A variety of feature ranking measures can be applied to imbalance datasets, including:

Information Gain (IG): This feature ranking measure estimates the decrease in entropy that occurs when the dataset is divided based on a specific attribute. It can help identify the features that are most useful for classification and is frequently used in decision trees.

ReliefF: ReliefF is a feature selection algorithm that weighs the differences between the closest examples of the same and other classes when determining the significance of each feature. It is helpful in identifying the features that most strongly separate the majority class from the minority class.

Gini Index: The Gini index calculates the likelihood that an instance would be erroneously classified based on the distribution of classes in a node. It can help identify the features that offer the most information for classification and is frequently used in decision trees.

Chi-squared test: This analysis determines if two categorical variables are independent of one another. It can be used to identify the features that significantly influence the class variable.

Mutual Information (MI): MI is a measure of how much information a feature imparts to the class variable. It can help identify the most informative features for classification and is frequently used in feature selection algorithms.

$$\text{Ent(I)} = -\sum_{x=0}^{m-1}\sum_{y=0}^{n-1} f(I((i,j)))\log_2 f(I((i,j))$$

$$\text{Energy} = \sum_{x=0}^{m-1}\sum_{y=0}^{n-1} f^2(I((i,j)))$$

The pseudo code provided outlines the steps for performing feature extraction from a dataset of liver images using optimized Principal Component Analysis (PCA). The input to the algorithm is the dataset X, which contains m instances (in this case, liver images) and p features (the pixel values of each image). The desired number of principal components k is also specified as an input. The output is a transformed matrix R, which is a reduced version of X containing only the most important features. The first step is to compute the mean vector u of the dataset X, which is the average value of each feature across all instances. This is done by summing all instances and dividing them by the total number of instances. Next, the covariance matrix C is computed for X, which measures how much each feature varies together with the other features across all instances. This is done by subtracting the mean vector from each instance, multiplying each instance by its transpose, and then summing across all instances. The eigenvalues λ and eigenvectors E of C are then computed using the eigenvalue method. Eigenvalues represent the variance explained by each eigenvector, and eigenvectors represent the direction of maximum variance in the dataset. The top k eigenvectors are then selected to form a new matrix E_k, which defines a new k-dimensional space in which to project the instances. The transformed matrix R is then computed by projecting X onto the k-dimensional space defined by E_k. This is done by multiplying X by E_k. The resulting matrix R contains the transformed instances with only the most important features. Finally, the algorithm returns R as the output. The use of optimized PCA for feature extraction can aid in the accurate diagnosis and treatment of liver diseases by identifying the most relevant features in large datasets of liver images.

2.3 Feature Ranking Algorithm

Probabilistic feature extraction measures are a type of feature extraction technique used in liver image segmentation processes. These measures rely on probabilistic models to extract features from liver images, which can be used to identify and segment the liver tissue.

There are several methods for probabilistic feature extraction in liver image segmentation, including the Hidden Markov Model (HMM), theta regulated Gaussian Mixture Models (TGMMs), and the theta control Markov Random Field (TMRF) model.

The HMM model represents the liver image as a series of hidden states and observed features, where the hidden states represent the underlying structure of the liver image, and the observed features represent the actual pixel values. The HMM model calculates the probability distribution of the hidden states given the observed features, which can be used to segment the liver tissue into different segments.

TGMMs represent the liver image as a combination of Gaussian distributions, with each distribution representing a particular type of liver tissue. The model calculates the probability distribution of the liver image given the observed features and uses this information to segment the liver tissue. The TMRF model represents the liver image as

a graph, where each pixel acts as a node, and the edges between nodes represent the spatial relationships between pixels. The model calculates the probability distribution of the liver image given the observed features by establishing a prior distribution and a likelihood function. The prior distribution captures the statistical characteristics of the liver image and the spatial layout of the liver tissue, while the likelihood function gauges how closely the observed features adhere to the prior distribution. By combining the prior distribution and likelihood function, the model can calculate the probability distribution of the liver image, which can be used to precisely segment the liver tissue.

2.4 K-joint Density Probabilistic based Segmentation Approach

The K-Density probabilistic segmentation approach estimates the probability density function of each liver tissue cluster using a kernel density estimator. This approach has several benefits when handling imbalanced liver datasets. It can handle liver images with a significant degree of class imbalance, where the minority class has a very low sample size and can withstand image noise and outliers. The K-Density probabilistic clustering approach can locate complex and non-linear liver tissue clusters, which is crucial for numerous real-world applications. One of the main benefits of the K-Density probabilistic clustering approach is its ability to be applied to both binary and multi-class liver segmentation issues. It can also be utilized for unsupervised learning when the number of liver tissue classes is unknown in advance. Another benefit of the K-Density probabilistic clustering approach is its ability to be used with other methods to enhance the performance of liver image segmentation models. Liver automatic segmentation using joint probability estimations involves finding the optimal segmentation of liver images by estimating the joint probability distribution of the data. The inputs to the algorithm include a filtered Imbalanced liver dataset and the number of segmentation regions K. The first step is to initialize the mixture model by randomly assigning parameters to K Gaussian mixture components. The Expectation step involves calculating the posterior probability of each region belonging to each of the K segmentation regions using Bayes' theorem. The Maximization step involves updating the parameters of each mixture component by calculating the mean, covariance matrix, and prior probability using the posterior probabilities calculated in the Expectation step. These steps are repeated until convergence is obtained, which is checked by monitoring the change in the log-likelihood of the data. If the log-likelihood does not change significantly, the algorithm stops and returns the final segmentation vector, which indicates the segmentation assignment for each region. The algorithm calculates the probability of selecting each segmentation region score based on its density using the joint probability distribution of the data. This approach can accurately segment the liver tissue even in the presence of imbalanced liver datasets and can locate complex and non-linear segmentation regions.

Automatic liver and tumor segmentation algorithm using joint probability estimations:
Inputs:
Dataset: filtered Imbalanced liver dataset of liver images
K: the number of segmentation regions
Outputs:
Segmentation: a vector of length N, where N is the number of regions in the dataset, indicating the segmentation assignment for each region
Steps:
Initialize the mixture model:
Randomly initialize the parameters for K Gaussian mixture components
Expectation step:
Calculate the posterior probability of each region belonging to each of the K segmentation regions using Bayes' theorem
Maximization step:
Update the parameters of each mixture component by calculating the mean, covariance matrix, and prior probability using the posterior probabilities calculated in the Expectation step
Repeat steps 2 and 3 until convergence:
Check for convergence by monitoring the change in the log-likelihood of the data
If the log-likelihood does not change significantly, stop the algorithm and return the segmentation
Return the final segmentation vector, which indicates the segmentation assignment for each region
Calculate the probability of selecting each segmentation region score based on its density using the joint probability distribution of the data as

$$P\left(z_i = k \mid x_i, \theta^{(t)}\right) = \frac{P\left(x_i \mid z_i = k, \theta^{(t)}\right) P\left(z_i = k \mid \theta^{(t)}\right)}{P\left(x_i \mid \theta^{(t)}\right)}$$

$$\gamma_{i,k}^{(t)} = P\left(z_i = k \mid x_i, \theta^{(t)}\right)$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_{i,k}^{(t)} x_i}{\sum_{i=1}^{n} \gamma_{i,k}^{(t)}}$$

$$\sigma_k^{2(t+1)} = \frac{\sum_{i=1}^{n} \gamma_{i,k}^{(t)} \left(x_i - \mu_k^{(t+1)}\right)^2}{\sum_{i-1}^{n} \gamma_{i,k}^{(t)}}$$

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_{i,k}^{(t)}}{n}$$

$$P(X \mid \theta) = P\left(z_i = k \mid x_i, \theta^{(t)}\right) \sum_{i=1}^{n} \log\left[\sum_{k=1}^{K} \pi_k^{(t+1)} \mathrm{N}\left(x_i \mid \mu_k^{(t+1)}, \sigma_k^{2(t+1)}\right)\right]$$

3. Ensemble Classification Framework

3.1 Non-linear SVM optimization with Kernel Function

The kernel function is composed of three distinct terms:

$$\ker_{\mathbf{n}}(d) = \underbrace{\exp^{-\sigma^2\left(d_1^2+\cdots+d_D^2\right)}}_{:=P(\mathbf{x})} \underbrace{\sqrt{\frac{\left(2\sigma^2\right)^{c_1+\cdots+c_D}}{c_1!\cdots c_D!}}}_{:=A(c)} \underbrace{d_1^{n_1}\cdots d_D^{c_D}}_{:=B(d,\mathbf{n})}$$

The first term, a(x), is a radial basis function (RBF) that measures the similarity between pairs of input vectors based on the distance from the origin. It decreases exponentially as the distance between the vectors increases, and its width is controlled by a parameter σ that balances the tradeoff between bias and variance in the SVM.

The second term, b(n), acts as a normalization constant and ensures that the kernel function is positive and finite. It is computed using the multinomial coefficient formula and is proportional to the number of possible monomials of degree n in D dimensions.

The third term, c(x,n), is a polynomial function of the input vector x and the degree vector n, which specifies the degree of each variable in the polynomial. It captures the non-linear interactions between the input features and enables the SVM to model complex decision boundaries.

3.2 Feature Ranking for Decision Tree Optimization Measure

The feature ranking measure for decision tree optimization is given by the following equations:

The feature ranking measure for decision tree optimization is given by the following equations:

FRM(Di) = Max{ρ1, ρ2}

where Di is the dataset, and ρ1 and ρ2 are calculated as follows:

ρ1 = (-CE(A2[]/A1[])^3) / ((sum(A1)[]_i)^3 * Corr(Di)^3)

ρ2 = (-CE(A1[]/A2[]) / ((sum(A2)[]_i)^3 * sqrt(Corr(Di)))

Here, CE(A2[]/A1[]) and CE(A1[]/A2[]) represent the conditional entropy of A2 given A1 and A1 given A2, respectively. Corr(Di) is the correlation coefficient of the feature Di with the class variable, and sum(A1)[]_i and sum(A2)[]_i represent the sums of the ith column of A1 and A2, respectively. N is the total number of observations, and m is the minimum of the number of rows and columns.

Additionally, a Max-Hellinger entropy-based ensemble learning model is proposed, where the PE (potential error reduction) value is calculated as follows:

PE = Math.cbrt(infogain(data)NHellinger(data)) * E(D) / chiValExp(data)

Here, data is the dataset, N is the total number of observations, E(D) is the entropy of the class variable, and chiValExp(data) is the chi-squared value of the data.

Finally, the proposed HER (Hellinger entropy ranking) measure is calculated as the maximum of the following three values:

$$PE = \text{Math.cbrt}\left(\inf \text{ogain}\left(\text{data}\right) * N * \text{Hellinger}\left(\text{data}\right)\right) * E\left(D\right) / \left(\text{chiValExp}\left(\text{data}\right)\right);$$

$$\text{ProposedHER} = \max\{\sqrt[3]{(\sum_{p=1}^{|D_p|}\sum_{n=1}^{|D_n|}(\sqrt[3]{D_p/|D_p|}-\sqrt[3]{D_n/|D_n|})^2)}, \text{corr}(D), PE\}$$

These rankings are then combined across all nodes to create a final ranking of the features.

3.3 Optimal KNN

The improved KNN involves computing the squared Euclidean distance between each instance in the dataset and a test sample. The distance is calculated using the formula (D (p1, p2)) = log(Σ(‖pi‖−‖pj‖)²), which is then used to sort the k-nearest neighbors of the test sample T according to their distances. The probability estimation for the sorted neighbors is then computed. For each instance ti in the k-neighbors (sort (k, D(T, p))), the distance probability is computed using the formula DistProb[] = (1 / √(2π)) ∫ e^(D(Ti, p)²) dD(Ti, p) / |N|, where N is the total attribute. This step is crucial in dealing with imbalanced liver datasets since it assigns weights to the neighbors based on their distance probabilities. Finally, for each test sample t in the k-neighbors (sort (k, D(T, p))), the class membership probabilities are computed, and the class is assigned to t using the classifier. This step involves computing the membership probabilities of each class and selecting the class with the highest probability.

3.4 Proposed U-Net Model

The Un-Net is a network architecture that builds upon the traditional U-Net model by making changes to the skip connection path, pooling path, and up-convolution path in the node structure. In the Un-Net, all output features in the node are connected to the next nodes and same-level encoder nodes, unlike in the conventional U-Net and most U-Net-based models where only the output features of the last convolution unit of the nodes are used as input for the next layers and the decoder node. The node structure in the Un-Net consists of n convolution units in each node, with subsequent units using dense connections that combine pooled features from upper nodes with previous convolution unit features. The transition node and decoder nodes are similarly structured, with the top decoder node connecting directly to the output. Additionally, the Un-Net implements deep supervision by multiple side-outputs fusion (MSOF) for better performance on image segmentation tasks.

4. Experimental Analysis

In this section, the proposed framework and its experimental findings are presented, evaluating the precision, accuracy, recall, and F-measure. Diverse imbalanced datasets were used for classification tasks, and liver segmentation stage. The learning rate was meticulously fine-tuned to 3e-5 for 30 epochs to enhance the training process. Over-fitting was prevented by utilizing a batch. The compact size of 8 and dropout rate of 0.2 make this model a valuable tool for medical image analysis. An early-stopping mechanism was also incorporated during the training phase to optimize performance. Two publicly available datasets, namely the LiTS and 3DIRCADb datasets, were utilized in this study. The LiTS dataset included 131 CT volumes for training and 70 for testing. However, adjustments were necessary to ensure accurate evaluation of the model's performance due to the absence of ground truth data for the testing subset. Ground truth data was manually curated by three experienced radiologists who meticulously collected CT scans. However, the dataset presented challenges due to variations in resolution and section spacing, ranging from 0.55mm to 1.0mm and 0.45mm to 6.0mm respectively, which added complexity to the analysis. Despite the challenges, the research was conducted meticulously, accounting for these variations to ensure the validity of the findings. The 3DIRCADb dataset, which consisted of 20 CT volumes, played a crucial role in the LiTS dataset, particularly volumes 28 to 47. The incorporation of this dataset improved the accuracy and comprehensiveness of the research, resulting in a more reliable outcome. To ensure a comprehensive evaluation of the network's performance, a rigorous stratified partitioning approach was employed, dividing both datasets into three distinct segments. This methodical approach allowed for a thorough assessment of the network's effectiveness, utilizing a training set of 90 CT scans (comprising of 85 LiTS volumes and 5 3DIRCADb volumes), a validation set of 11 LiTS volumes, and a testing set of 30 CT scans (comprising of 15 LiTS volumes and 15 3DIRCADb volumes) carefully selected through a meticulous curation process. The proposed model is compared with existing models such as FCNN (Deep learning fully CNN), MC-FCNN (Multichannel Fully CNN), and U-Net on different liver tumor regions.

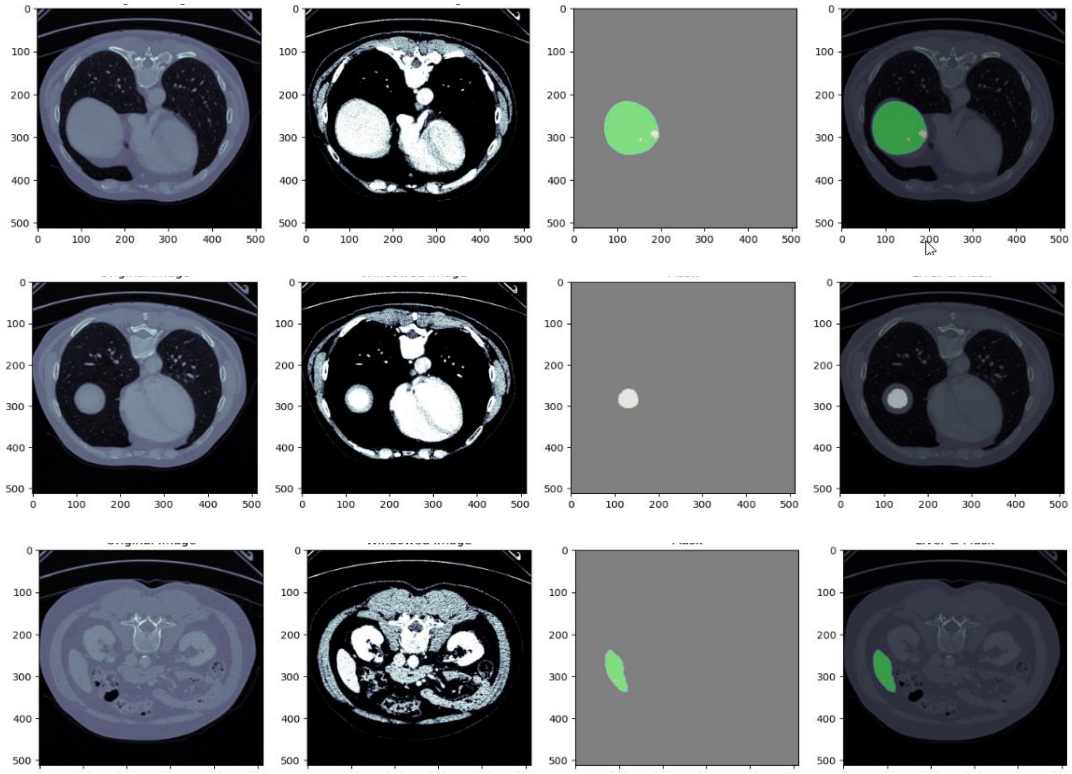


Fig. 3. Liver Tumor Segmentation and Detection

Figure 3 describes each liver tumor image and its corresponding tumor regions. Initially, a liver image with heterogeneous noisy tumors is taken as input for segmentation and tumor detection. The proposed segmentation-based classification framework is used to detect the tumor regions.

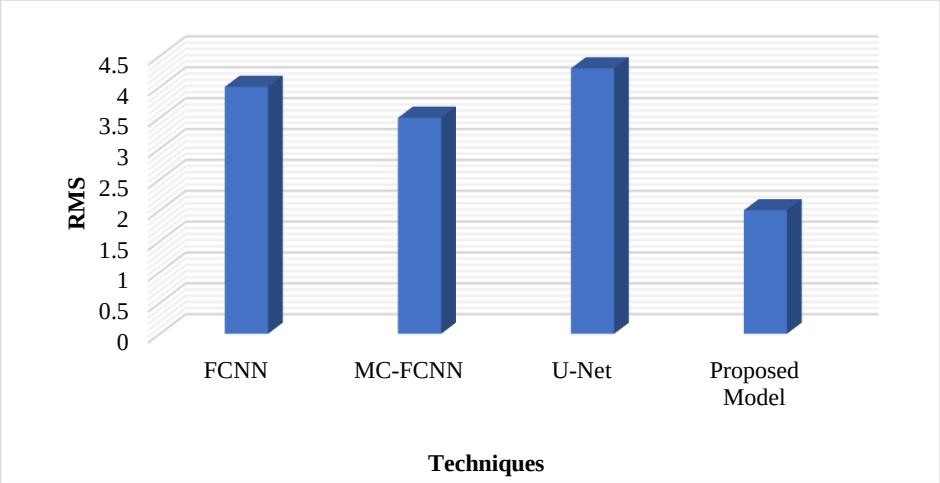


Fig. 4. Comparative analysis of proposed segmentation-based classification model RMS to the conventional models RMS on liver tumor dataset

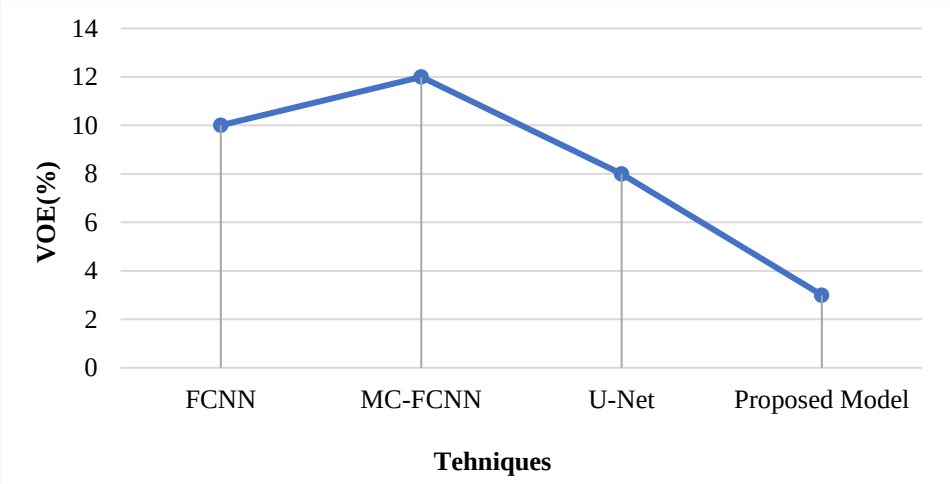


Fig. 5. Comparative analysis of proposed segmentation-based classification model RMS to the conventional models VOE (%) on liver tumor dataset

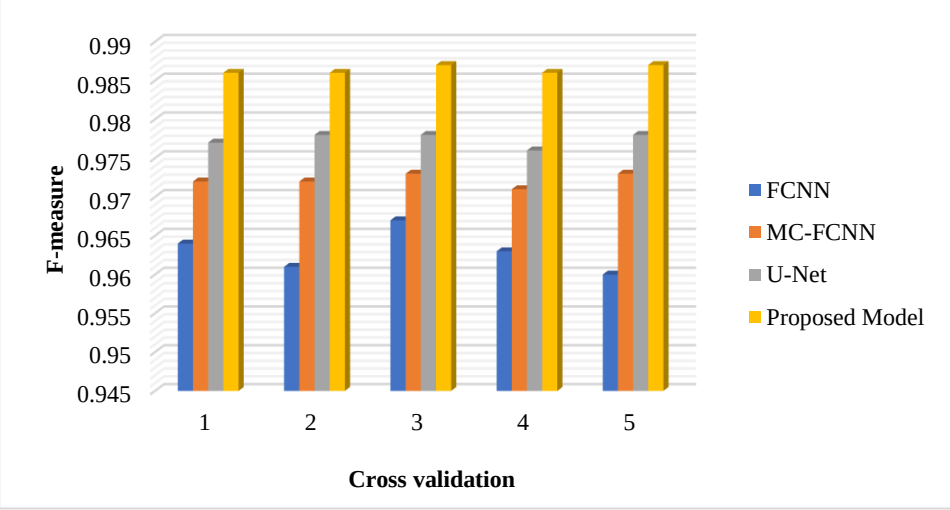


Fig. 6. Comparative analysis of proposed liver segmentation-based classification approach and existing models using F-measure for noisy tumor detection on different heterogeneous images

Table 1. Comparative analysis of proposed liver segmentation-based classification approach and existing models using accuracy for noisy tumor detection on different heterogeneous images

CV	FCNN	MC-FCNN	U-Net	Proposed Model
#1	0.962	0.974	0.976	0.982
#2	0.966	0.972	0.976	0.981
#3	0.965	0.971	0.975	0.988
#4	0.967	0.971	0.977	0.985
#5	0.965	0.971	0.977	0.986

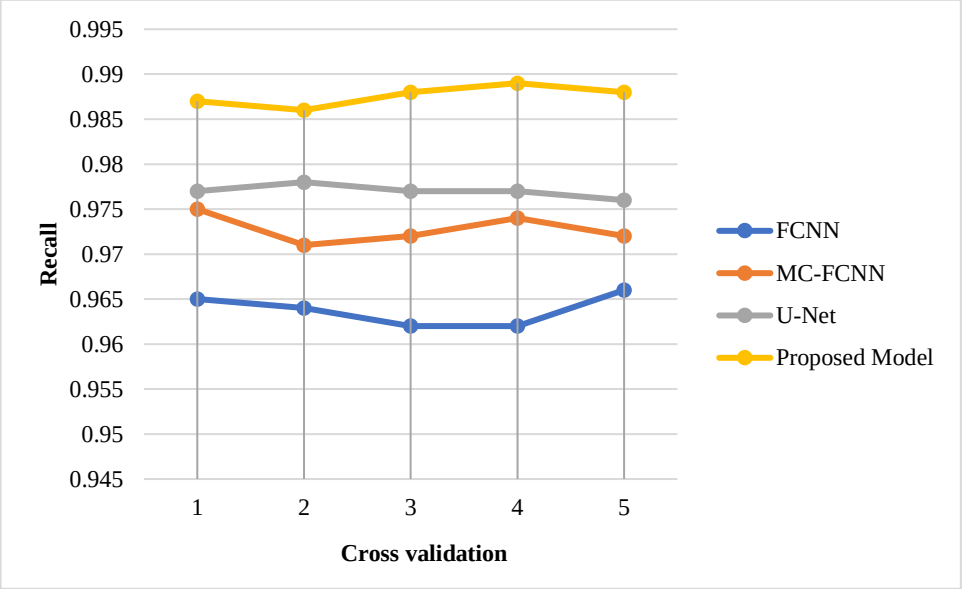


Fig. 7. Comparative analysis of proposed liver segmentation-based classification approach and existing models using recall for noisy tumor detection on different heterogeneous

Table 2. Comparative analysis of proposed liver segmentation-based classification approach and existing models using AUC for noisy tumor detection on different heterogeneous

FCNN	MC-FCNN	U-Net	Proposed Model
0.964	0.974	0.977	0.984
0.962	0.972	0.977	0.984
0.96	0.975	0.978	0.985
0.963	0.972	0.976	0.989
0.962	0.974	0.978	0.989

Table 3. Comparative analysis of proposed liver segmentation-based classification approach and existing models using accuracy for noisy tumor detection on different homogenous and heterogenous tumor dataset.

FCNN	MC-FCNN	U-Net	Proposed Model
0.96	0.971	0.977	0.989
0.961	0.973	0.976	0.982
0.967	0.973	0.976	0.987
0.962	0.973	0.978	0.984
0.966	0.974	0.975	0.983
0.966	0.973	0.976	0.983
0.965	0.971	0.978	0.98
0.966	0.973	0.978	0.983

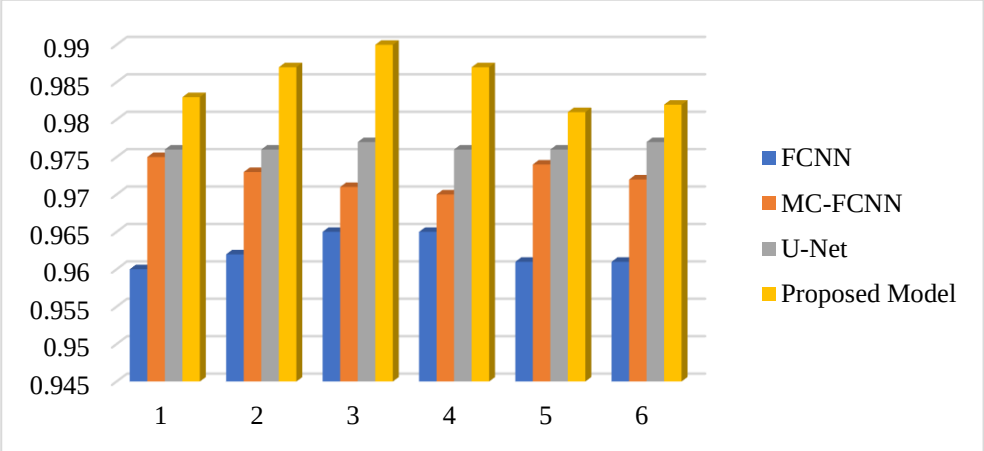
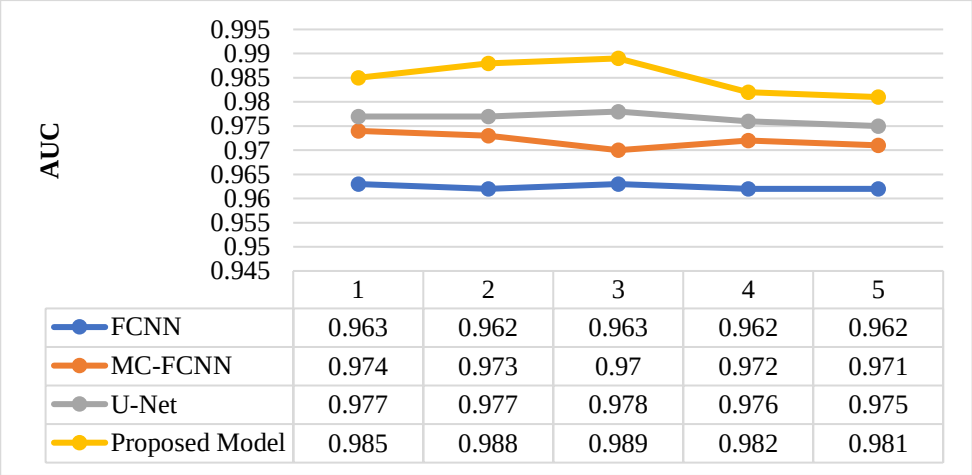


Fig. 8. Comparative analysis of proposed liver segmentation-based classification approach and existing models using recall for noisy tumor detection on different homogenous and heterogenous tumor dataset.



	1	2	3	4	5
FCNN	0.963	0.962	0.963	0.962	0.962
MC-FCNN	0.974	0.973	0.97	0.972	0.971
U-Net	0.977	0.977	0.978	0.976	0.975
Proposed Model	0.985	0.988	0.989	0.982	0.981

Fig. 9. Comparative analysis of proposed liver segmentation-based classification approach and existing models using AUC for noisy tumor detection on different homogenous and heterogeneous tumor dataset

5. Conclusion

In this paper, a novel k-joint probabilistic based multi-tumor classification model is proposed on different tumor imbalance regions. To ensure high-quality predictions on imbalanced liver datasets, the study proposes an optimized ensemble classification model that utilizes optimized filtering and classification approaches. Additionally, a novel strategy is proposed to handle missing data, imbalanced liver classes, feature selection, and ensemble classification approaches to improve the true positive rate and error rate on imbalance databases. However, 3D convolutions come with high computational costs, and 2D convolutions have limited spatial information utilization. The quality of input data, including missing feature values, feature noise, and imbalanced liver classes, significantly impacts the efficiency of classification approaches, making it necessary to ensure high-quality input data to achieve optimal results. This work proposes an optimized ensemble classification model based on k-joint probabilistic segmentation to address the challenges of both homogeneous and heterogeneous liver tumor detection. Furthermore, novel approaches for image filtering, feature extraction, and ranking are proposed to enhance the classification process for imbalanced liver tumor regions. The experimental results indicate that the proposed classification model outperforms existing models in terms of accuracy, recall, precision, and AUC.

References

[1] P. Lv, J. Wang, and H. Wang, “2.5D lightweight RIU-Net for automatic liver and tumor segmentation from CT,” Biomedical Signal Processing and Control, vol. 75, p. 103567, May 2022, doi: 10.1016/j.bspc.2022.103567.

[2] Q. Zhang, Y. Liang, Y. Zhang, Z. Tao, R. Li, and H. Bi, “A comparative study of attention mechanism based deep learning methods for bladder tumor segmentation,” International Journal of Medical Informatics, vol. 171, p. 104984, Mar. 2023, doi: 10.1016/j.ijmedinf.2023.104984.

[3] R. Rong et al., “A Deep Learning Approach for Histology-Based Nucleus Segmentation and Tumor Microenvironment Characterization,” Modern Pathology, p. 100196, Apr. 2023, doi: 10.1016/j.modpat.2023.100196.

[4] Y. Chen et al., “A deep residual attention-based U-Net with a biplane joint method for liver segmentation from CT scans,” Computers in Biology and Medicine, vol. 152, p. 106421, Jan. 2023, doi: 10.1016/j.compbiomed.2022.106421.

[5] G. Tong and H. Jiang, “A hard segmentation network guided by soft segmentation for tumor segmentation on PET/CT images,” Biomedical Signal Processing and Control, vol. 85, p. 104918, Aug. 2023, doi: 10.1016/j.bspc.2023.104918.

[6] M. O. Khairandish, M. Sharma, V. Jain, J. M. Chatterjee, and N. Z. Jhanjhi, “A Hybrid CNN-SVM Threshold Segmentation Approach for Tumor Detection and Classification of MRI Brain Images,” IRBM, vol. 43, no. 4, pp. 290–299, Aug. 2022, doi: 10.1016/j.irbm.2021.06.003.

[7] O. Alpar, “A mathematical fuzzy fusion framework for whole tumor segmentation in multimodal MRI using Nakagami imaging,” Expert Systems with Applications, vol. 216, p. 119462, Apr. 2023, doi: 10.1016/j.eswa.2022.119462.

[8] Z. Diao, H. Jiang, and T. Shi, “A unified uncertainty network for tumor segmentation using uncertainty cross entropy loss and prototype similarity,” Knowledge-Based Systems, vol. 246, p. 108739, Jun. 2022, doi: 10.1016/j.knosys.2022.108739.

[9] G. Chen et al., “An improved 3D KiU-Net for segmentation of liver tumor,” Computers in Biology and Medicine, p. 107006, May 2023, doi: 10.1016/j.compbiomed.2023.107006.

[10] J. Zhang, H. Jiang, and T. Shi, “ASE-Net: A tumor segmentation method based on image pseudo enhancement and adaptive-scale attention supervision module,” Computers in Biology and Medicine, vol. 152, p. 106363, Jan. 2023, doi: 10.1016/j.compbiomed.2022.106363.

[11] R. Ranjbarzadeh and S. B. Saadi, “Automated liver and tumor segmentation based on concave and convex points using fuzzy c-means and mean shift clustering,” Measurement, vol. 150, p. 107086, Jan. 2020, doi: 10.1016/j.measurement.2019.107086.

[12] G. Z. Ferl et al., “Automated segmentation of lungs and lung tumors in mouse micro-CT scans,” iScience, vol. 25, no. 12, p. 105712, Dec. 2022, doi: 10.1016/j.isci.2022.105712.

[13] R. R. Savjani, M. Lauria, S. Bose, J. Deng, Y. Yuan, and V. Andrearczyk, “Automated Tumor Segmentation in Radiotherapy,” Seminars in Radiation Oncology, vol. 32, no. 4, pp. 319–329, Oct. 2022, doi: 10.1016/j.semradonc.2022.06.002.

[14] R. V. Manjunath and K. Kwadiki, “Automatic liver and tumour segmentation from CT images using Deep learning algorithm,” Results in Control and Optimization, vol. 6, p. 100087, Mar. 2022, doi: 10.1016/j.rico.2021.100087.

[15] A. Qayyum, A. Lalande, and F. Meriaudeau, “Automatic segmentation of tumors and affected organs in the abdomen using a 3D hybrid model for computed tomography imaging,” Computers in Biology and Medicine, vol. 127, p. 104097, Dec. 2020, doi: 10.1016/j.compbiomed.2020.104097.

[16] Y. Ren, D. Zou, W. Xu, X. Zhao, W. Lu, and X. He, “Bimodal segmentation and classification of endoscopic ultrasonography images for solid pancreatic tumor,” Biomedical Signal Processing and Control, vol. 83, p. 104591, May 2023, doi: 10.1016/j.bspc.2023.104591.

[17] R. Vankdothu and M. A. Hameed, “Brain tumor MRI images identification and classification based on the recurrent convolutional neural network,” Measurement: Sensors, vol. 24, p. 100412, Dec. 2022, doi: 10.1016/j.measen.2022.100412.

[18] Daher, M. G., Trabelsi, Y., Ahmed, N. M., Prajapati, Y. K., Sorathiya, V., Ahammad, S. H., ... & Rashed, A. N. Z. (2022). Detection of basal cancer cells using photodetector based on a novel surface plasmon resonance nanostructure employing perovskite layer with an ultra high sensitivity. Plasmonics, 17(6), 2365-2373.

[19] Reddy, A. P. C., Kumar, M. S., Krishna, B. M., Inthiyaz, S., & Ahammad, S. H. (2019). Physical unclonable function based design for customized digital logic circuit. International Journal of Advanced Science and Technology, 28(8), 206-221.

[20] Ü. Budak, Y. Guo, E. Tanyildizi, and A. Şengür, “Cascaded deep convolutional encoder-decoder neural networks for efficient liver tumor segmentation,” Medical Hypotheses, vol. 134, p. 109431, Jan. 2020, doi: 10.1016/j.mehy.2019.109431.

[21] X. Wang, S. Wang, Z. Zhang, X. Yin, T. Wang, and N. Li, “CPAD-Net: Contextual parallel attention and dilated network for liver tumor segmentation,” Biomedical Signal Processing and Control, vol. 79, p. 104258, Jan. 2023, doi: 10.1016/j.bspc.2022.104258.

[22] Zuhayer, A., Abd-Elnaby, M., Ahammad, S. H., Eid, M. M., Sorathiya, V., & Rashed, A. N. Z. (2022). A Gold-Plated Twin Core D-Formed Photonic Crystal Fiber (PCF) for Ultrahigh Sensitive Applications Based on Surface Plasmon Resonance (SPR) Approach. Plasmonics, 17(5), 2089-2101.