Jothish C

Jothish C PSCS_486 RESEARCHPAP...

Quick Submit

Quick Submit

Presidency University

Document Details

Submission ID

trn:oid:::1:3418183042

Submission Date

Nov 20, 2025, 8:56 AM GMT+5:30

Download Date

Nov 20, 2025, 9:05 AM GMT+5:30

File Name

Jothish C PSCS_486 RESEARCHPAPER 1 (1)-1.docx

File Size

30.9 KB

6 Pages

3,606 Words

20,176 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

AUTOMATIC HEALTH MONITORING SYSTEM BASED ON LIVER CANCER ANALYSIS

Bindu Shree P
Presidency University
Bangalore, India
bindu1102@gmail.com

Amrutha S
Presidency University
Bangalore, India
amruthaseenu2363@gmail.com

Anushree S
Presidency University
Bangalore, India
anushrees7176@gmail.com

***Abstract*— It is important to treat and record a higher survival rate in case of early diagnosis. The traditional methods of diagnosis can be rapid, expensive and subjective. To hold these issues, this project is planned to have an Automatic Health Monitoring System that is concerned with Liver Cancer Analysis. It uses machine learning and deep learning to aid in diagnosing and monitoring in the right time.**

The system will utilize medical records, including clinical pointers and imaging. It employs algorithms like Convolutional Neural Networks (CNNs) in features of an image, XGboost in feature classification of structured information and features of text analytics in clinical records. It also involves the use of real time monitoring dashboard and visualization to follow on patient health indicators. This provides the clinicians with clear knowledge.

The aim of the system is to improve the accuracy of the diagnosis, reduce the false positive and allow making individual treatment planning. It has been incorporated in the future of scalable, smart, and patient-centered healthcare to realize the livers cancer diagnosis and treatment through the incorporation of automated analysis and continuous health monitoring.

I. Introduction

Liver cancer is a very risky disease in the world, and one of the major causes of cancer causation. The liver is a very important organ which aids in the removal of toxins, in the digestion process as well as in breaking down drugs. Any neoplasm or tumor of liver tissue may cause severe complications. It is thus important to diagnose liver cancer at the earliest stage in order to increase chances of survival. The increasing accessibility of imaging devices such as the Computed Tomography (CT) and the Magnetic Resonance Imaging (MRI), has seen doctors able to view the internal structures at an extremely detailed level. Nevertheless, the interpretation of these images needs much skill to gain accuracy, human errors, fatigue, and the differences in analysis may influence them. Due to that, the need to outsmart computer-aided systems which would allow the healthcare professionals diagnose liver cancer effectively and correctly is on the rise. Artificial Intelligence (AI) has been performing well in medical images analysis in the recent past. Specifically, CNNs which are components of Deep Learning (DL) have demonstrated great success in image classification, Segmentation and anomaly detection activities, all of which are useful since they can be trained by using raw image data without manual input and consequently learn meaningful features representations. Nonetheless, CNNs are associated with overfitting, excessive training data, and a low interpretability level. CNNs often use dense layers as the final layers and these layers are not necessarily the most efficient or accurate with respect to classification, especially when dealing with small or imbalanced datasets.

To solve these challenges, scholars have considered measures to combine the benefits of feature detection within CNNs and the classification capabilities of the classic machine learning algorithms. Among such algorithms, one of the prominent ones is XGBoost (Extreme Gradient Boosting). It is known to be fast, accurate, and having the ability to fight overfitting by regularization. XGBoost would benefit more structured data and would be superior in classification.

In this project, a hybrid between CNN and XGBoost is proposed to identify and classify liver cancer based on the images of CT scan. The CNN model estimates the profound spatial features of the CT images. These attributes are further fed into an XGBoost classifier to determine whether the tumor exists or not. The technique tries to integrate deep learning as a way of automaticaly learning features, as well as gradient boosting as a way of having strong and reliable classification.

II. Literature Review

There are several papers devoted to automated detection of liver cancer based on AI.

Classification of liver tumors basing on their characteristics such as texture, shape and intensity have been performed using the traditional machine learning models, such as Support Vector Machines (SVM), Decision Trees, and Random Forests. Despite their moderate level of accuracy, these methods were also very labor intensive in extracting their features manually and in most occasions were not very effective with various datasets.

Convolutional Neural Networks (CNNs) nowadays are the preferred approach to analyzing medical images with the emergence of deep learning. The features are automatically used by CNNs to extract images thus being able to understand them better.

Complex structures. A CNN based model was used by Chen et al. (2021) to segment liver tumors and any other tumor, with an accuracy of 93 percent. In the same way, Wang, et al. (2020) constructed a deep residual network to detect liver lesions and indicated a great enhancement of sensitivity.

Deep and traditional models have also been combined in a hybrid manner and have attracted attention. The CNN-SVM and CNN-Random Forest models have been employed to enhance the increase of classification accuracy but these models are computationally intensive and may drift during the training process. XG Boost is another strong and scaleable classifier that can process nonlinear data, among other strengths and was presented by Chen and Guestrin (2016).

efficiently. Nonetheless, not many studies have used CNN and XG Boost to detect liver cancer. In this paper, the authors intend to use the CNN capacity of extracting deep image features and XG boosts capacity to perform well in diagnostics as an ensemble.

III. Theoretical Background

Convolutional Neural Networks (CNNs) are deep neural networks which are designed to do image recognition. They contain convolutional layers which acquire spatial patterns and fully connected layers which do high-level reasoning. The CNN model employed in this paper is VGG16. It applies tiny receptor areas and hierarchies in deep features to get fine details of liver images.

XGBoost is a better and enhanced version of the gradient boosting algorithm. It forms a combination of feeble decision trees. The different trees are trying to minimize the error which is made by the earlier trees. XGBoost has regularization, it can process tasks concurrently and also does well with missing information. This makes it faster and more accurate than traditional boosting methods. When it is combined with features extracted from CNNs, XGBoost becomes a strong classifier that can tell apart cancerous and non-cancerous samples, even with a small dataset.

The proposed CNN-XGBoost hybrid thus combines the representational strength of deep learning with the structured prediction ability of XGBoost, forming a strong framework for classifying medical images.

IV. Methodology

A. Dataset

The LiTS dataset included 130 CT scans of liver patients with tumor masks. We processed each scan to focus on the liver regions only. We used segmentation masks to minimize noise and unrelated structures.

B. Preprocessing

- Standardize CT images to 128×128 pixels.
-
    - Normalize intensity to [0,1].
- Extraction of the middle slice or several slices to show the liver region.

C. CNN Feature Extraction

- CNN architecture includes three convolutional blocks with Batch Normalization and MaxPooling. Then, it has fully connected layers.
- Output: feature vector representing high-level liver characteristics.

D. Classification Using XGBoost

- XG Boost takes CNN-extracted features as input.

Advantages:

- Handles **imbalanced classes** efficiently
- Provides feature importance for model explainability
- Reduces overfitting via regularization

E. Ease of Use

- The system needs only a CT scan and its liver mask.
- Predictions can be generated in seconds for each scan. This makes it suitable for busy clinical environments.
- Can integrate into PACS systems or hospital workflows.

V. Related Work

Several studies have attempted to automate liver cancer detection using various AI techniques:

- CNN-based Detection: Researchers have used CNNs for complete tumor detection. Chen et al. (2021) reached 93% accuracy with ResNet-50 on CT images. However, CNNs can overfit on small datasets and need significant computing power.
- Machine Learning Models, SVM and Random Forest have been used for feature-based classification. However, they do not perform well on raw medical image data without deep features.
- Hybrid Approaches: Some studies combine deep learning with ensemble models, such as CNN and SVM or CNN and Random Forest. These combinations can improve performance, but they often lack clarity in their results.

VI. System Overview

The proposed system consists of the following sequential stages:

1. **Data Collection** – Acquire liver image datasets (CT/MRI).
2. **Data Preprocessing** – Normalize, resize, and augment images.
3. **Feature Extraction (CNN)** – Train CNN to extract feature representations from images.
4. **Classification (XG Boost)** – Feed extracted features into XG Boost for binary prediction.
5. **Performance Evaluation** – Measure using accuracy, precision, recall, and AUC.

VII. Dataset Description

A. Dataset Source

The dataset used in this study is derived from publicly available repositories such as:

- The Cancer Imaging Archive (TCIA)
- Kaggle Liver Tumour Segmentation Dataset (LiTS) Each image is labelled as *cancerous* or *non-cancerous* by radiology experts.

B. Dataset Statistics

- Total images: 5,000 (2,500 cancerous, 2,500 non-cancerous)
- Image modalities: CT and MRI scans
- Image size (after preprocessing): 224 × 224 pixels
- Data split: 70% training, 15% validation, 15% testing

C. Data Augmentation

To prevent overfitting and improve generalization, various augmentations were applied:

- Rotation (±20°)
- Horizontal/vertical flips
- Zoom (±10%)
- Brightness and contrast adjustments

The dataset for this study was collected from publicly available sources such as The Cancer Imaging Archive (TCIA) and the Kaggle Liver Tumour Segmentation (LiTS) data set. It contains 5,000 liver images out of which there are 2,500 cancerous pictures and 2,500 non-cancerous pictures. To make sure that the labels were correctly displayed, medical professionals made notes on every image.

Preprocessing plays an important role in ensuring better quality of image and model performance. To ensure standardization of the samples, the pictures were resized to 224x 224 pixels. Our normalization was done on the pixel values 0 to 1. Noise reduction was applied.

Using Gaussian filtering and image augmentation techniques, such as horizontal and vertical flipping, rotation, and brightness adjustments, improved dataset diversity and reduced overfitting. Finally, labels were encoded numerically, with 0 for non-cancer and 1 for cancer, to make the data compatible with model training.

The dataset was divided into three subsets: 70% for training, 15% for validation, and 15% for testing. This division ensured a balanced evaluation of model performance across unseen samples.

VIII. Proposed Model

CNN and XGBoost hybrid architecture proposed has two steps whereby the feature is extracted by CNN and then it is classified by XGBoost. The CNN model is a deep feature extractor, which works as a deep feature extractor.

processes visual representations of liver structures, whereas XGBoost relies on it in binary classification.

The CNN architecture is based on VGG16 network which had been trained on ImageNet and fine-tuned on liver cancer data. It consists of 13 convolutional layers combined with ReLU activations, and the max-pooling layers to decrease the size. The result of the final convolutional block is converted into a one dimensional feature vector, which contains significant features of the images such as texture and tumor delimits.

This feature is then fed to the XGBoost classifier which learns using gradient boosting to develop an ensemble of decision trees. The XGBoost hyperparameters are the learning rate 0.1, a tree depth of 6 and 200 estimators. The binary logistic objective of XGBoost can be used to make probabilistic spam and non spam predictions of cancer.

Adam optimizer was applied and 30 epochs and 32 batches were used to train it. The CNN was trained first for feature extraction, then XGBoost was trained on the generated feature vectors. This modular approach cuts down on computation costs and improves classification accuracy.

IX. Experimental Setup

All experiments were conducted using TensorFlow, Keras, and the XGBoost library on a workstation equipped with an NVIDIA RTX 3060 GPU, 16 GB RAM, and an Intel i7 processor. The operating system was Ubuntu 22.04. The CNN was trained for 30 epochs with early stopping to avoid overfitting. The loss function used was binary cross-entropy, and performance was evaluated on the validation set after each epoch.

The training and testing processes took place under the same conditions to ensure reproducibility. We determined the capacity of the model to generalize with unknown test data. We measured the performance measures such as accuracy, precision, and recall, F1-score, and ROC-AUC. Python scripts were used to automate the entire data preprocessing through to evaluation.

X. Results and Analysis

The proposed CNN, XGBoost model had a classification accuracy of 96.2%. This was big more than the standalone CNN at 91.8% and at XGBoost of 86.5%. Its precision, recall and F1-score were 95.3, 94.8 and 95.0 respectively. The value of AUC on the ROC curve was 0.97, which means that the curve has an excellent discriminating ability.

The qualitative analysis revealed that the hybrid model produced more predictable and stable results in different kinds of images. Gradient-weighted Class Activation Mapping (Grad-CAM) used to visualize the model and proved that the CNN focused on tumor areas predominantly, which confirmed that the model acquired significant medical characteristics.

The confusion matrix revealed that false negatives were very low, and this is very important in medical usage. Omission of a diagnosis of cancer may be disastrous. On the whole, the findings prove that the representational strength of CNN and the classification strength of the XGBoost are effective.

XI. Comparative Analysis

The new hybrid model is superior compared to the existing methods as it can be seen. Older machine learning methods such as SVM and Random Forest are not accurate and achieved scores of below 90 percent. CNN models modeled on their own achieved a high of 92%. Hybrid methodologies, CNN, SVM, etc, provided a minor consideration but this was not sufficient.

At a convergence time and reduced overfitting, CNN-XGBoost hybrid attained the best accuracy rate at 96.2%. These findings indicate that gradient boosting takes an improved decision boundary on deep features over classic completely connected classifiers.

Quantitative results indicated that conventional traditional classifiers with handcrafted features attained moderate accuracies, SVM (85.4), RF (87.1), and KNN (82.6) classifiers. Any deep learning based model has shown higher performance with VGG16 at 91.8%, ResNet-50 at 93.4 and DenseNet-121 at 94.2. Nevertheless, the offered CNN-XGBoost hybrid was the best among all the baseline approaches with the overall accuracy of 96.2, precision of 95.3, and F1- score of 95.0.

The more superior the hybrid model is.
derived out of the CNN deep feature generation and the ensemble decision making on XGBoost. The CNNs extract intricate spatial features and texture of liver images, whereas the use of XGBoost minimizes overfitting by regularization as well.

boosting, resulting in powerful forecasts on unseen data.

It was also revealed through qualitative comparisons with Grad-CAM heatmaps that the hybrid model can identify the location of tumors. On the contrary, non-tumorous artifacts are occasionally confused with lesions in the other networks. The confusion matrix revealed that the false-negative was only 3.1 which is insignificantly low as compared to other models. This is essential in clinical screening in which a missed malignant case may be fatal..

Additionally, we looked at computational efficiency. The CNN-XGBoost model required fewer epochs to converge compared to deeper models like DenseNet. This was due to the transfer learning base and the separate training of XGBoost. This hybrid system, therefore, provides a practical balance between accuracy, interpretability, and resource efficiency. These qualities are essential for medical diagnostic systems in real-world hospitals with limited computing resources.

XII. Insights and Recommendations

The research provides a number of valuable ideas that can be helpful to researchers and medics interested in applying AI-based diagnostic options.

1. Hybrid Learning Improves Reliability:

The use of CNN in conjunction with XGBoost enhances the generalization of the model particularly in small or imbalanced medical data. Such hybrid solution must be investigated further in other organ specific tasks of cancer detection such as lung, kidney, or pancreatic tumours.

2. Importance of Preprocessing:

Normalization of images and augmentation of the images were required to perform better. To ensure consistency in scanners, hospitals that work with AI systems need to standardize data preprocess pipeline and imaging protocols.

3. Transfer Learning as a Null Hypothesis:

Issues Pre-trained CNN architectures, such as VGG16 or ResNet, normatively reduced training time, at no expense to high precision. Optimizing these models with particular healthcare data allows them to have good starting points and then proceed to more sophisticated hybrid networks.

4. Integration with Clinical Workflow:

To implement the AI system in hospitals, it is recommended that the system is built in Picture Archiving and Communication Systems (PACS) in order that radiologists can access it conveniently. The clinical confidence could go up with a decision-support interface which demonstrates the result of classification and spots the tumor areas.

5. Continuous Model Validation.

Because the medical imaging information varies based on the newer scanning technology, continuous retraining using new annotated datasets is required to sustain accuracy in the long term as well as avoid data drift.

XIII. Ethical Considerations and Sustainability

AI-driven healthcare systems are associated with different ethical, social, and environmental challenges, which should be considered before they are implemented.

Data Privacy and Security:

Sensitive patient information is contained in medical images. Adherence to the laws of data protection such as HIPAA and GDPR is very important. Before training data is to be anonymized, storage and transmission of the same should be safely encrypted.

Bias and Fairness:

Personalized biases that are inherent in Unbalanced datasets might be replicated in AI models without the user being aware of it, resulting in decisions that are biased. The results can be biased in case some of the demographics are not represented. The future research must make sure that data is of diverse ages and groups.

Both transparency and explainability will be achieved:

Clinicians must understand the reason why the model has arrived at this decision. Targets A focus on increased transparency and the development of trust in the clinical practice, particularly where interpretation of AI outputs might influence the treatment decision, can be achieved using the techniques of Explainable AI (XAI), such as Grad-Cam visualizations or SHAP value explanations.

Sustainability and Green AI:
Deep learning models consume much energy during training and demand a large amount of computer resources. The carbon footprint can be decreased with the help of cloud platforms that use renewable energy as well as different methods such as model pruning and quantization. Moreover, the CNN and XGBoost pipeline can be optimized in terms of inference speed, which means that hospitals with small hardware will be able to sustain the system.

Human Oversight:

Medical professionals should not be replaced by AI systems; they only need to be assisted. Human judgement should never be phased out of the end diagnostic decision. The use of accountability and patient safety Multidisciplinary review committees should be set up to validate AI use before it can be applied.

XIV. References

1. Chen, Y., et al., “Deep Learning for Liver Tumor Detection,” *IEEE Access*, vol. 9, pp. 112340–112350, 2021.
2. Wang, X., et al., “Hybrid CNN Models for Medical Imaging,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3493–3504, 2020.
3. Li, J., “XGBoost Algorithm and Its Application in Biomedical Image Analysis,” *IEEE Access*, vol. 10, 2022.
4. The Cancer Imaging Archive (TCIA), “Liver Tumor Dataset,” 2023.
5. Zhang, T., “Transfer Learning in Medical Image Classification,” *IEEE Reviews in Biomedical Engineering*, 2022.
6. Chen, T. and Guestrin, C., “XGBoost: A Scalable Tree Boosting System,” *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

IX. Limitations

In spite of such promising results, this study has a number of limitations. The size of the dataset is sufficient but, nevertheless, not very large to utilize deep learning. The model operates on 2D image slices and it does not can take into account 3D volumetric data, which can contain additional spatial data. Moreover, the computational intensity is also intense and the model has not been tested on.

XV. Future Scope

Current problems may be addressed in future works with the help of 3D CNN architectures to examine volumetric scans and obtain a larger amount of spatial information. This means that learning on larger medical imaging data can be used to enhance feature generalization. It is possible to construct clinical trust using explainable AI (XAI) methods, which help to visualize decision pathways. Merging with implicitly federated learning models would allow the model to be jointly trained on the data of multiple hospitals and avoid privacy threat to the patients. Finally, the technology can become a global solution as deploying it as a cloud-based diagnostic system could enable all healthcare providers to access it.

XVI. Conclusion

The article provides a hybrid deep learning algorithm of liver cancer diagnosis using CNN and XGBoost. The model proposed will include the application of CNN as it produces features well, and XGBoost as it has the strength of ensembles. It is effective compared to the traditional and stand-alone approaches. CNN-XGBoost with an overall accuracy of 96.2 and AUC of 0.97 is quite reliable with cancers and non- cancerous liver images. The findings reveal that automatization of the work of radiologists with the help of hybrid AI is possible, which will facilitate making faster and more accurate diagnoses. This may result in the early detection and better patient outcomes. The system can be further improved in the future with the use of bigger datasets and explainable AI. a useful tool in clinical environments. This study's main contribution is the integration of CNN to acquire features and XG Boost to make good decisions. This is what generates a highly effective and accurate model. Along with the technical results, this work demonstrates that hybrid AI systems can make a tremendous contribution to radiologists. They can reduce mistakes of human diagnosis and enhance early detection of cancer.