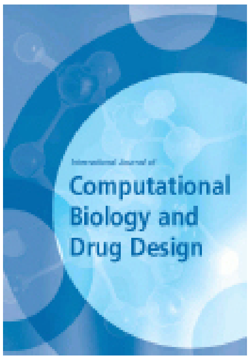


International Journal of Computational Biology and Drug Design

ISSN online: 1756-0764 - ISSN print: 1756-0756
https://www.inderscience.com/ijcbdd

A review on speech organ diseases and cancer detection using artificial intelligence

M. Swathi, Rajeshkannan Regunathan, Suresh Kumar Nagarajan

DOI: 10.1504/IJCBDD.2023.10054029

Article History:

Received:	16 August 2022
Last revised:	07 November 2022
Accepted:	11 November 2022
Published online:	17 April 2023

A review on speech organ diseases and cancer detection using artificial intelligence

M. Swathi, Rajeshkannan Regunathan*
and Suresh Kumar Nagarajan*

School of Computer Science and Engineering,
Vellore Institute of Technology,
Vellore, 632014, Tamil Nadu, India
Email: swathi.madiraju46@gmail.com
Email: rajeshkannan.r@vit.ac.in
Email: sureshkumarnagarajanvit@gmail.com
*Corresponding authors

Abstract: Cancer is an abnormal expansion of tissues and is among the common illnesses in India, accounting for 0.3 fatalities each year. It can appear in any way and is incredibly challenging to spot in its early phases. Thus, Speech organ-related cancer detection using image processing-based techniques and ML is a challenging field in the medical domain, which involves early detection and diagnosis of cancer. This research defines the methods, algorithms, and datasets used by the existing researchers on speech organ diseases. The results from the state-of-the-art works are evaluated by accuracy, false-positive rate, and area under the ROC curve (AUC). The merits and demerits of these approaches are examined, which paves the way for future research in reducing the death rate of patients. The literature studies motivate us to develop early detection of cancer of all speech organs with resources of mobile-related applications to use in real-time will be our future vision. Thus, 60–80% of all speech-related organ infections or cancerous cases can be detected at early stages by this one mobile-related application, which will be beneficial for people in reducing the death rate of patients.

Keywords: deep learning; larynx; pharynx; throat cancer; tonsillitis; oral cavity; voice disorders; ROC curve; speech organs.

Reference to this paper should be made as follows: Swathi, M., Regunathan, R. and Nagarajan, S.K., (2023) ‘A review on speech organ diseases and cancer detection using artificial intelligence’, *Int. J. Computational Biology and Drug Design*, Vol. 15, No. 4, pp.336–356.

Biographical notes: M. Swathi completed her MTech from G. Pulla Reddy Engg College from Kurnool, Andhra Pradesh. She has been pursuing research as a full-time research scholar in the department of CSE, Vellore Institute of Technology, Vellore, since December 2020. She has published three conferences in Springer proceedings. Her research interests include machine learning, image processing, and deep learning.

Rajeshkannan Regunathan has received his MTech in Computer Science and Engineering from SASTRA University, Tanjore, and a PhD in Computer

Science and Engineering from Vellore Institute of Technology, Vellore, India. He works as an Associate Professor at VIT, Vellore, India. His area of interest is cloud computing, artificial intelligence, natural language processing and data science. He has 15 + years of teaching experience. He has published more than 35 papers in international journals and conferences. He is a Member of CSI, IEEE (WIE).

Suresh Kumar Nagarajan has completed his Master of Engineering from Anna University, Chennai, Tamil Nadu. He did his doctoral studies on Genetic based classification Algorithms for combined LANDSAT and ENVISAT Images, Vellore Institute of Technology, Vellore, India. He has 20 years of teaching experience. He has more than 54 publications in national, international journals and conferences. His research interests include image processing, computational intelligence and big data.

1 Introduction

Cancer is a deadly disease worldwide that can affect any part of the body leading to the abnormal growth of cells and spreading to other parts. World Health Organization (WHO) statistics reported around 9 million deaths worldwide due to cancer in 2018–2021; in 2020–2021, deaths increased to 10.5 million. Among that, some of the most affected parts by cancers in 2020-2021 are the lung (around 2 million new cases), and death caused by these cancers increases yearly (Abujamous et al., 2018).

Based on the statistics of oral cancer and other speech-related cancer, more than 10 billion people are suffering from different stages of infections or cancer without detection at early stages. Thus, the speech-related medical domain is a promising field for research, especially in helping people with one mobile application and saving/alerting the person at early stages by spreading awareness and precautions in real-time.

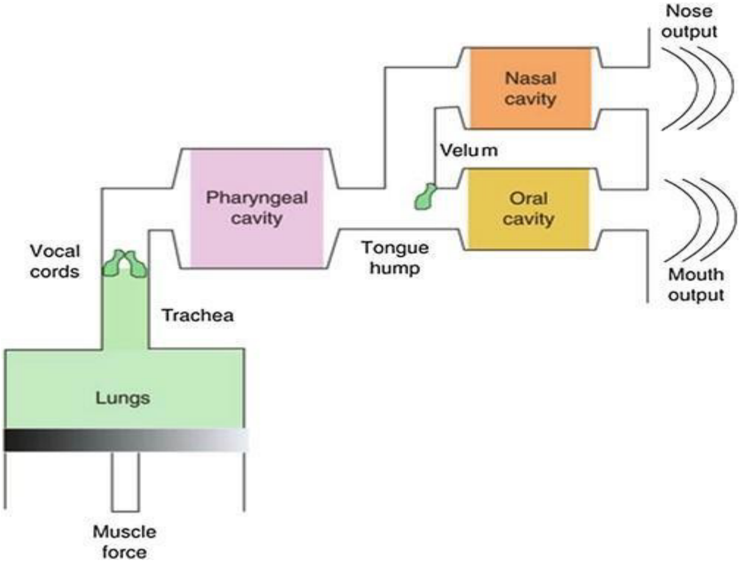
Speech creation is a complex human technique to communicate and convey emotions. The lungs, vocal tract, and lips are all part of the human speech production system, as shown in Figure 1.

The vocal tract (the space between the lips and the cords) functions resonator, spectrally shaping the periodic input (much like a wind instrument) (Rabiner and Juang, 1993). The larynx, Pharynx, and oral cavity are all human organs for producing speech, as shown in Figure 1.

If any organs impact diagnosis procedures, it will take longer. Thus, we focus on setting up a system to identify and recognise abnormalities with speech organs (larynx, Pharynx, and oral cavity). Some of the significant risk aspects of cancers are pollution, tobacco and alcohol usage, unhealthy diet, and physical slothfulness, affecting the human body and leading to cancer tumours.

The diseases and other related issues of speech organs (larynx, Pharynx, and oral cavity) with specific statistical data are in Table 1.

Figure 1 Speech production mechanism (see online version for colours)



Source: Rabiner and Juang (1993)

Table 1 Summary of the diseases of speech organs

S. no	*Speech organ*	*Definition*	*Statistics*
1	Pharyngitis (Faden et al., 2016; Follmann et al., 2016)	Pharyngitis refers to the medical term for a sore throat	In 2010, 1.814 million people visited the emergency room with Pharyngitis, with 692,000 children under 15
2	Larynx Cancer (Nocini et al., 2020)	Laryngitis is an inflammation of your voice box (larynx) from overuse, irritation, or infection	Laryngeal cancer incidence, prevalence, and mortality are predicted to be 2.76, 14.33 cases, and 1.66 deaths per 100,000 people, respectively, resulting in 3.28 million deaths yearly
3	Oral Cavity Diseases (Ferlay et al., 2010)	The oral cavity is bounded by the teeth, tongue, hard palate, and soft palate. The most common locations for cancer in the oral cavity are tongue	Oral cancer is the sixth most common cancer worldwide. It is one of the ten most incessant diseases worldwide, and its rate of occurrence is increasing every decade
4	Mouth Cancer (Ferlay et al., 2010)	Mouth cancer refers to a variety of cancers that begin in the mouth. These are typically prevalent on the lips, tongue, and hard palate, but they can also begin in the cheekbones, mucilage, floor of the mouth, tonsils, and salivary glands	The men in the 70 to 75-year age group (64.8%) and western regions of India (58.4%) in the 60–69-year age group

Table 1 Summary of the diseases of speech organs (continued)

S. no	*Speech organ*	*Definition*	*Statistics*
5	Tongue Cancer (Sankaranarayanan et al., 1998)	Tongue cancer is a type of tumour that develops in the tongue's cells. The tongue can be affected by various cancers, and it begins in the skinny, flat squamous cells covering the exterior of the tongue	From 60 to 69-year age group in the northern part of the world had the highest likelihood of acquiring tongue cancer (58.4%), followed by males in the 70–75-year age group (37.2%)
6	Lip Cancer (Sankaranarayanan et al., 2005)	Lip cancer can develop in any area of the top or lower lip; however, the lower lip is the most commonly affected. Lip cancer is a kind of mouth cancer (oral cancer). Most lip cancers are malignant tumours that	Males in the northeast area in the 70-year age group had the highest Automatic Anatomy Recognition (AAR) (37.1%), followed by males in the western regions in the 70- to 75-year age group who had the lowest AAR (12.3%)

Based on the above-listed statistics of cancer in speech-related organs, which majorly affects speech production, it is necessary to predict cancer tumours.

1.1 Motivation

The early prediction comparatively reduces the severity and economic cost as the treatment for early diagnosis is less. Early detection of cancer includes being conscious of cancer symptoms and following proper advice from the doctors. Early detection of speech organ-related cancer can reduce risk prominence as a leading cause of voice loss or death.

1.2 Scope and objectives

The survey's primary goal is to present an extensive on speech organ infection/cancer detection using machine learning (ML) and deep learning (DL) approaches. Furthermore, the advantage and disadvantages of state-of-the-artworks are discussed and reviewed.

1.3 Survey organisation

The survey is organised as follows, and Section 2 provides the background of the ML-based algorithms with appropriate diagrams. Section3 studies existing papers about cancer detection using image processing-based ML methods with suitable diagrams and tables. Section 4 provides the methodology of the proposed work. Section6 provides the discussion, followed by a conclusion in Section 7 respectively.

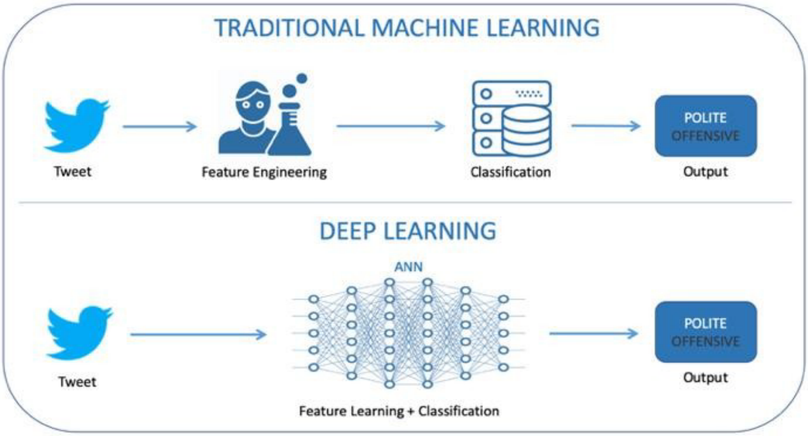
1.4 ML and DL algorithms

This section's survey of various ML-based approaches in speech organ disease and cancer detection provides basic preliminary knowledge, including clear explanations and suitable diagrams.

1.5 Machine learning and its types

Machine learning (ML) and deep learning (DL) are the subset of Artificial intelligence(AI) that allows machines to mimic a human brain to solve any real-time applications problems like detection, classification, etc., intelligently by past information (Pahwa and Agarwal, 2019; Bulbul and Ünsal, 2010). Figure 2 represents the applications of ML.

Figure 2 ML and DL approaches algorithms (see online version for colours)



Some of the significant ML and DL algorithms are:

1.5.1 Support vector machine (SVM)

The hyperplane separates the data points based on the highest distance between them. The kernels are responsible for selecting the optimal hyperplane that can separate the two classes of data points based on the highest distance, which can be formulated as,

$$W_{eigT} IP_{vx} + fav = 0 \tag{1}$$

The above equation $WeigT$ represents the weight function for the vector input IP_{vx} and fav the good points based on the high distance (Parveen and Singh, 2015). Adding a kernel, which uses more features in the SVM, leads to increased accuracy in classification. Figure 3 represents the SVM model.

1.5.2 K nearest neighbours (KNN)

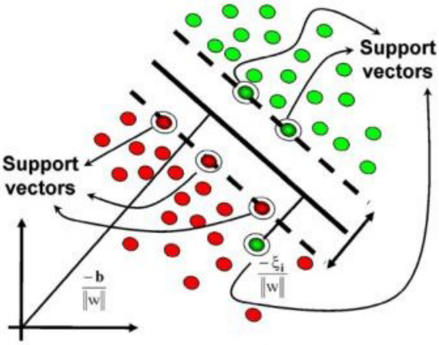
The KNN is a simple supervised learning classification algorithm. The sample to be classified is labelled based on the uncategorised data closest to it (Sun and Huang, 2010). The input is assigned to each class to represent the nearest data points of the classes. The distance measurement of data points is done by using Euclidean distance. The distance measurement by each measurement technique can be formulated as,

$$EUC^D = \left(\sum_{m=1}^{K} (h_m - r_m)^2 \right)^{\frac{1}{2}} \tag{2}$$

$$HAM^D = \sum_{m=1}^{K} |(h_m - r_m)^2| \tag{3}$$

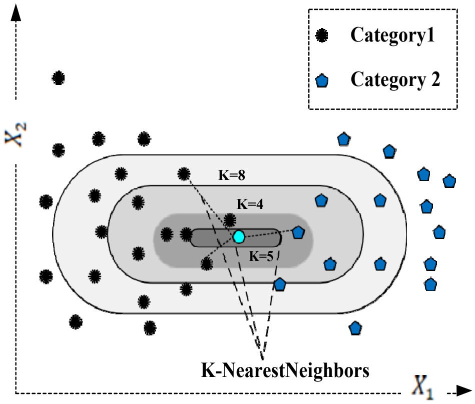
The above two equations represent the measurement technique using Euclidean distance (EUC^D) and hamming distance (HAM^D), respectively. Figure 4 depicts the KNN model.

Figure 3 SVM (see online version for colours)



Source: Parveen, and Singh (2015)

Figure 4 Illustration of the KNN model (see online version for colours)



Source: Sun and Huang (2010)

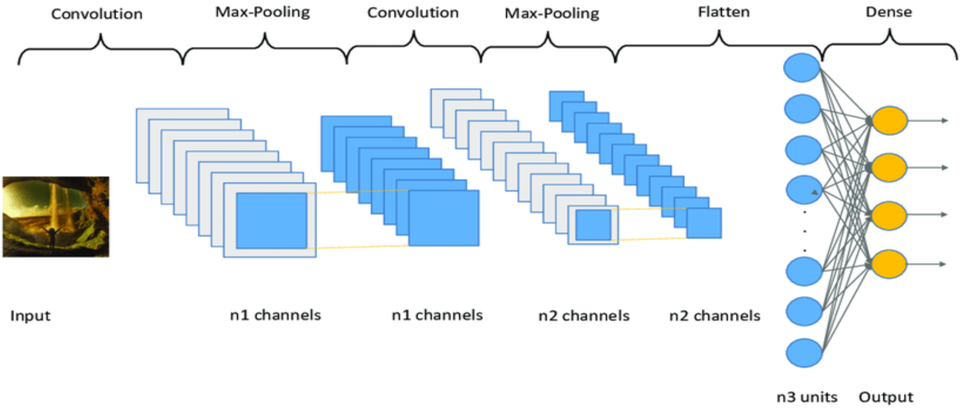
1.5.3 Neural networks

The neural network contains neurons, weight, and activation function (Mangileva et al., 2020), as shown in Figure 5. The computation of the input function is done by using weight and activation functions to transform the results by element-wise non-linearity function, which can be formulated as,

$$Ip = \alpha(wei_{ght} Ip + Fav) \tag{4}$$

The above equation represents that functions and results compute input function (*Ip*) are validated using the activation function $\alpha()$. The mathematical methods in neural networks tune the weights. In the training stage, raw data is learned by the model, which contains the error and learning rate function (Parpulov et al., 2018; Subramaniam et al., 2019). The layers in CNN are convolutional, max pool, and fully connected. These layers are effectively extracted features with fewer computational resources. Figure 5 represents the neural network model.

Figure 5 Neural network (see online version for colours)



Source: Mangileva et al. (2020)

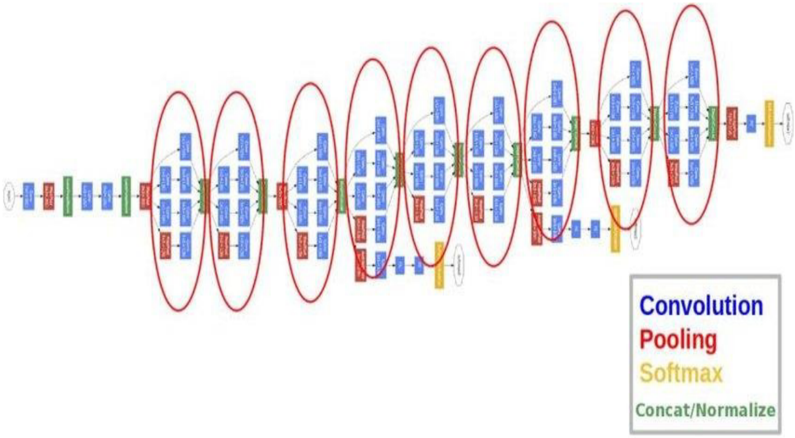
1.5.4 GoogleNet

The 22-layer network architecture GoogLeNet (Binti Mat Kasim et al., 2018) is shown in Figure 6. Unlike AlexNet, the design comprises stacked building pieces of fully connected layers. The GoogLeNet architecture was chosen because of its depth. The convolution and pooling layers' links are blue and red, respectively. During learning, the architecture utilises a softmax classifier for effective erroneous back-propagation. The softmax activation function used in this GoogleNet architecture is shown in yellow. The connection between the layers is established in green colour. Only the primary classifier is employed during implication. This classifier receives input from a single, linked layer, which analyses the highest-level building block information. The activations of the two highest-level building blocks are concatenated in the proposed GoogLeNet++ architecture before being sent to the fully linked layer (Binti Mat Kasim et al., 2018).

2 Existing studies of cancer detection on speech-related organs

This section briefly discusses the literature survey on speech organ-related diseases using ML DL algorithms. The machine learning and deep learning algorithms are in detail in Section 2. The commonly used speech-related organ datasets for cancer detection using image-based, machine-learning, or deep-learning algorithms are mentioned in the following subsections.

Figure 6 Google Net architecture (see online version for colours)



Source: Binti Mat Kasim et al. (2018)

2.1 Datasets used in existing studies

The datasets detect cancer or disease in speech-related organs like the Pharynx, larynx, oral cavity, or voice disorders. In our survey, most of the researchers have used standard datasets like Mendeley Data repositories, video laryngoscopy (Yoo et al., 2020), Saarbruecken Voice Database (SVD) (Kim et al., 2020), Single Nucleotide Polymorphisms (SNP) (Xiong et al., 2019), Massachusetts Eye and Ear Infirmary Database (Mohammed et al., 2020). In addition, some researchers utilised a private dataset created in partnership with local hospitals.

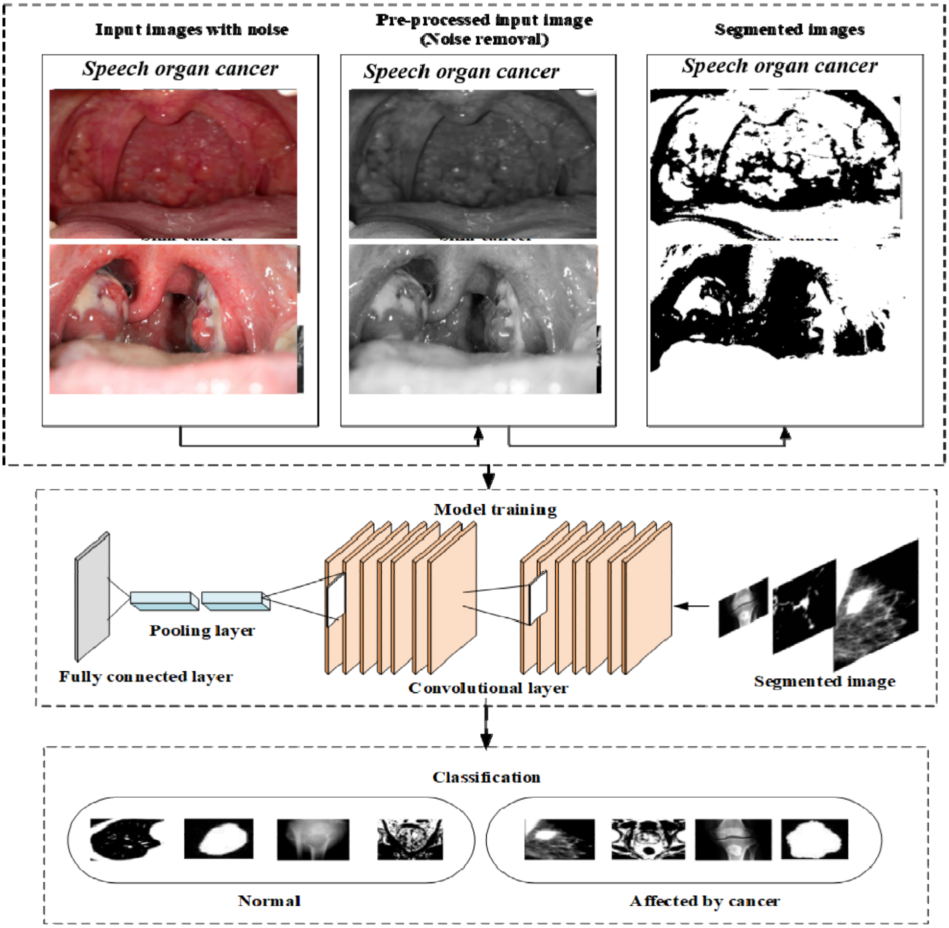
2.2 Literature survey

The speech production mechanism in humans, as discussed in Section 1, Figure 1, helps us to understand speech pathology, in which speech organs are required to produce sound. Out of which, a billion people suffer from different infection levels and cancer, primarily affecting speech. Thus, Extensive literature studies were carried out on speech-related organs, especially the larynx, Pharynx, and oral cavity, to detect cancer/infection using advanced machine learning, deep learning algorithms, and other image-based techniques.

In Askarian et al. (2019), the idea is that they manipulated images by transforming and correcting colour using the related algorithms and classifying streptococcal Pharyngitis from healthy throat using ML techniques. They used an intelligent phone camera to capture images. Here they used the YCbCr range as feature extraction. In Yoo (2020) developed a model using deep learning techniques to quickly identify highly intense Pharyngitis with a more timely diagnosis. Here they used RGB as feature extraction. Fujimura et al. (2022) developed a computer-assisted method using a deep learning framework to classify voice quality through the auditory evaluation scale. Here they used automatic feature extraction. Finally, Fang et al. (2019) proposed a comprehensive voice detection system that integrates normalised MFCC features and a DNN classifier. Here they used Mel frequency cepstral coefficients (MFCCs) as feature extraction. The DNN's performance is far better than traditional SVM and GMM in improvising detection accuracy dependent on three responsible features using both MEEI and FEMH databases.

It covers an advanced positive time-frequency analysis method with accurate zero, first, and second-order moments integrated with machine learning to detect abnormal voice signals (Georgopoulos, 2020). MFCC and Wavelets Modulation Spectrum were employed in this study. It is a state-of-the-art infrastructure technique to detect and categorise oral cancer from hyperspectral imaging of the examining maxillofacial region, according to Jeyaraj et al. (2020). They extracted features based on size, orientation, and chemical information in this case. It recognises sinusitis photos using medical information and a neural network (Jirawanitcharoen et al., 2009) (NN). Kim et al. (2021) is a CNN-based automated laryngeal mass detection algorithm and an embedded pilot system for home-based self-screening, as shown in Figure 7. Figure 7 shows the schematic diagram for cancer detection of the larynx using advanced machine learning, i.e., CNN. Phensadsaeng and Chamnongthai (2017) Proposes designing and implementing an automatic tonsillitis monitoring and detection system on the ISE utilising the VHDL programming language. Size and colour were considered features.

Figure 7 Cancer detection using CNN algorithms (see online version for colours)



It was reviewed in Akshara and Latchoumi (2021) the ML and DL-based research initiatives to classify carcinoma of the throat. In this case, they employed a boxplot and principal component analysis (PCA). In BenAicha and Ezzine (2016) focuses on building a feature extraction for identifying and categorising laryngeal cancer by exploring distinct glottal flow metrics. The temporal and frequency characteristics are determined once the glottal flow is removed. Automated speech detection based on machine learning and deep learning algorithms enable noninvasive yet objective detection of vocal changes in laryngeal cancer with accuracy levels that may exceed human performance, according to (Kim et al., 2020).

They employed MFCC feature extraction in this study (s). This study (Inaba et al., 2020) built and tested a diagnostic AI model for SLPC. They employed a single-shot multibox detector (SSD), and You Only Look Once (YOLO) for feature extraction. We can develop a computer-aided diagnosis (CAD) system using machine learning with low-level pre-processing and feature extraction, as demonstrated by Singh and Maurya (2021). They employed two different features: texture-based LBP descriptors and GLCM features. Sharma (2014) uses a genetic database to create multilayer perceptron (MLP) and support vector machine models (SVM) for predicting the survivorship of oral cancer patients who visit the ENT OPD.

According to Nayak et al. (2006), the two approaches of PCA and ANN achieve excellent specificity and sensitivity in objective discriminating among premalignant, normal, and malignant oral tissues. Sharma and Om (2015) aims to create a data mining model for the early identification and prevention of oral cancer utilising probabilistic neural networks and general regression neural networks (PNN/GRNN). The feature selection was made with WEKA3.7.9.

Using the gravitational search optimised echo state neural networks (GSOESNN) technique, Al-Maaitah and AlZubi (2018) analyses oral X-Ray pictures and detects oral cancer. Gray level cooccurrence matrix is used in this application (GLCM). The AI-based diagnostic system (Tamashiro, 2020), which used a CNN, displayed great diagnosis accuracy, particularly with NBI. Mitsuhiro Kono's study demonstrated that developing an AI system for pharyngeal lesion identification is promising, with good sensitivity and adequate specificity (Kono, 2020).

The summary of the extensive research studies on speech-related diseases with description, disease, and organ affected, and the algorithm used is discussed in Table 2.

Table 2 Review of various techniques on speech organ diseases

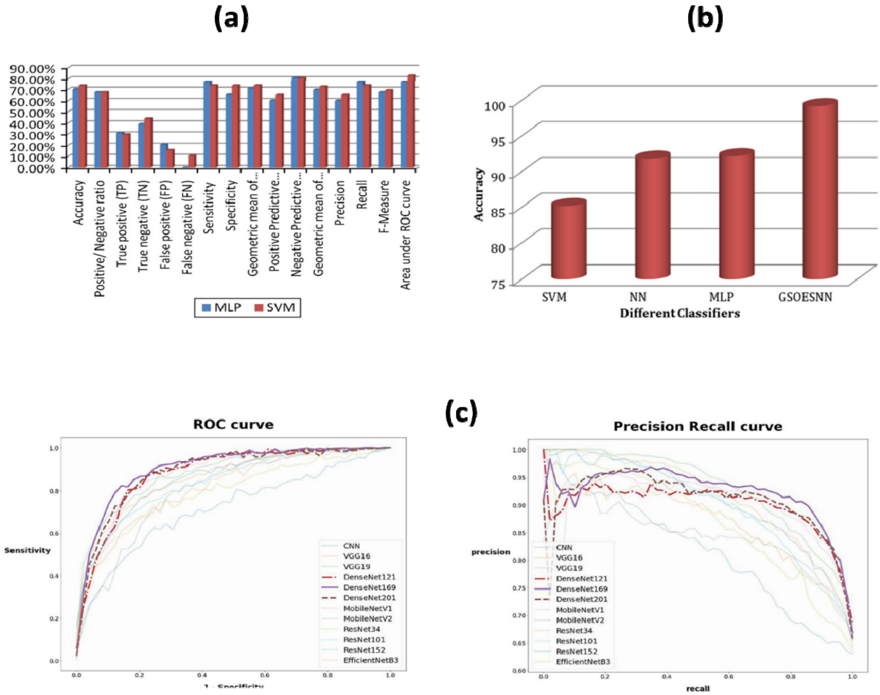
S. no.	*References*	*Speech organ*	*Disease*	*Algorithm*	*Performance*
1	Askarian et al. (2019)	Throat	Streptococcal Pharyngitis	K-NN, Colour intensity	Accuracy = 93.9%, Sensitivity = 87.5%, Specificity = 88%
2	Yoo et al. (2020)	Throat	Severe Pharyngitis	Cycle GAN model, CNN, deep learning	Accuracy = 95.3%, Sensivity = 95%, Specificity = 95%
3	Fujimura et al. (2022)	Throat	Voice disorders	1D-CNN model	Accuracy = 77.1%

Table 2 Review of various techniques on speech organ diseases (continued)

S. no.	*References*	*Speech organ*	*Disease*	*Algorithm*	*Performance*
4	Fang et al. (2018)	Throat	Voice disorders	Deep neural network (DNN), support vector machine, and Gaussian mixture model	Accuracy = 99.32%
5	Georgopoulos (2020)	Throat	Voice disorders	GoogLeNet	–
6	Jeyaraj et al. (2020)	Throat and oral cancer	Cancer	SVM, SVM-PCA, DBM models	Accuracy of 91.55% and 94.75%
7	Jirawanitcharoen et al. (2009)	Throat	Tonsils	Artificial neural network (ANN)	Correction = 90%
8	Kim et al. (2021)	Throat	Laryngeal	CNN	Accuracy = 80%
9	Phensadsaeng and Chamnongthai (2017)	Throat	Tonsillitis	Fuzzy logic	Accuracy = 96.4%
10	Akshara and Latchoumi (2021)	Throat	Cancer	Supervised classification algorithm	Accuracy = 98.55%
11	BenAicha and Ezzine (2016)	Larynx	Cancer	ANN	Accuracy = 96.9%
12	Kim et al. (2020)	Larynx	Cancer	1D-CNN	Accuracy = 85%, Sensitivity = 78%, Specificity = 93 %
13	Inaba et al. (2020)	Larynx	Laryngopharyngeal cancer	RetinaNet for object-detection	Accuracy = 97.3%
14	Singh and Maurya (2021)	Larynx	Cancer	SVM, NN, NavieBaseian	–
15	Sharma and Om (2014)	Oral cavity	Cancer	ANN, SVM, Multilayer perceptron	SVM Accuracy = 73.56%
16	Sharma and Om (2015)	Oral cavity	Cancer	ML (PCA), Regression Neural Network	Sensitivity = 0.6%, Specificity = 1.0%; Sensitivity = 1.0%, Specificity = 1.0%
17	Al-Ma'aitah and AlZubi (2018)	Oralcavity	Cancer	Gravitational search optimised echo state neural networks (GSOESNN) approach	Accuracy = 99.2%
18	Tamashiro et al. (2020)	Pharynx	Cancer	CNN	Accuracy = 73.5%
19	Kono (2020)	Pharynx	Pharyngeal Cancer	CNN	Accuracy = 66%

In comparison to MLP, SVM exhibits greater accuracy in accuracy, true negative, false negative, specificity, geometric mean of sensitivity and specificity, positive predictive value, geometric mean of PPV and NPV, precision, F-measure, and area under the ROC curve. Figure 8(a) compares the performance of MLP and SVM models based on several criteria.

Figure 8 Classification metric curves: (a) oral cavity cancer (Sharma and Om, 2014); (b) oral cavity cancer (Al-Maaitah and AlZubi, 2018) and (c) tongue cancer (Heo et al., 2022) (see online version for colours)



In contrast to other methods like support vector machine (89.2%), Neural Networks (94.1%), and Multi-layer Perceptron (95.2%), GSOESNN method consumes a high oral cancer recognition rate (99.2%) as in Figure 8(b). This is due to the method's efficient use of extracted features, neural network parameters, and updating procedures. Therefore, the GSOESNN accurately and efficiently detect oral cancer.

In this investigation, six distinct models – CNN, ResNet, EfficientNet, VGGNet, MobileNet, and DenseNet – were applied. The CNN model, the most fundamental image classification model, served as a benchmark for assessing other models' performance. Each model can perform better at prediction when the layers are deeper. VGGNet, ResNet, and DenseNet were models that shared a massive skeleton. Using these related models, we could identify trends in the data and the most relevant model. This is shown in Figure 8(c).

MobileNet and VGGNet are used to quickly check the findings by adding logic to locate data features more effectively. They have reasonably quick learning rates, meet the required performance, and have fast learning rates. Since ResNet, DenseNet, and

EfficientNet are made up of many deep layers, their learning rate is somewhat slow, but their performance is adequate. In comparison to ResNet, DenseNet performs better and uses fewer parameters. When moving across layers, ResNet combines features by summing them; however, DenseNet is different since it concatenates the features rather than adding them.

An extensive literature survey was conducted from 2006 to 2020 on the larynx, pharynx, oral cavity, and throat cancers. However, very little research was carried out on this cancer disease detection in speech-related organs. Out of which, more papers are on the larynx and oral cavity to detect cancer/infection with an accuracy of ~80–95% for ML and DL algorithms, especially ANN, SVM, and CNN. In addition, a few works were done on throat-related datasets, mainly used for identifying voice disorders, tonsils, and cancer using advanced ML algorithms like DNN, SVM, ANN, and CNN with an accuracy of ~73–93%. On the other hand, little research was done on developing cancer detection by developing a mobile-based application, VHDL, and VLSI technologies.

Based on an understanding of the literature studies, most of the work is done for cancer detection on different speech-related organs, especially oral cancer and the larynx. The merits of the approaches described above are that they have used machine learning and deep learning algorithms to detect cancer for each speech organ with good accuracy and other evaluation metrics. As per examining these studies, the demerits of these approaches pave the way for future research in reducing the death rate of patients with one mobile application. Thus, this motivates us to develop a mobile application that detects the cancer of all speech-related organs. One more challenging aspect here is incorporating the images of the larynx, Pharynx, oral cavity, tongue, and lips to train the model. Inserting the camera to capture the images of speech organs feeds into the application, showing the results as either cancer or non-cancer.

2.3 Speech-related cancer detection using ML/DL

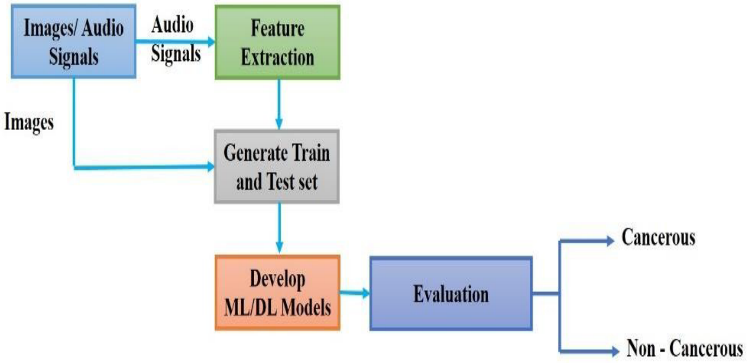
In this section, the survey of various ML-based approaches in cancer detection is provided extensively. Generally, the image processing-based cancer detection approaches perform a series of processes which are as follows,

- image acquisition
- feature extraction
- generate train and test set
- classification (ML/DL) models
- evaluation metrics.

Figure 9 represents the overall process flow of ML-based cancer detection. This section comprising of significant research works in the respective fields.

The steps for detecting or classifying the cancerous or non-cancerous Pharynx, larynx, oral cavity images, or voice signals using ML/DL algorithms are given below. Before using an ML algorithm, the images of speech organs must be pre-processed and turned into a set of features. Figure 9 depicts the technical flow of detecting cancer in speech-related organs using algorithms.

Figure 9 Process flow to detect cancer using ML algorithms (see online version for colours)

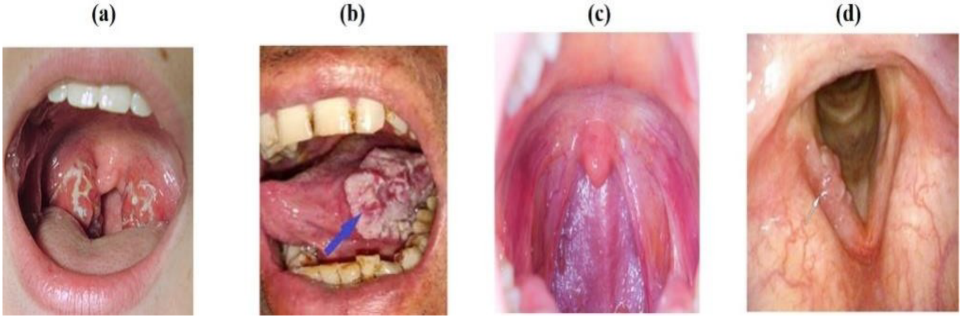


Generally, the image processing-based cancer detection approaches perform a series of processes which are as follows:

2.3.1 *Image acquisition (Images)*

Images or audio signals are considered as input to detect cancerous or non-cancerous. Experts must manually categorise a given set of speech data in audio files or images as reflecting cancerous or non-cancerous. Images should either be Pharynx, tonsillitis, or larynx. A few of the input pharynx or larynx images are shown in Figure 10. If the input is images, then we feed the input image dataset directly in Step 3.

Figure 10 Cancerous: (a) tonsillitis (Yoo et al., 2020); (b) oral (Jeyaraj et al., 2020); (c) throat (Askarian et al., 2019) and (d) laryngeal images (Ben Aicha and Ezzine, 2016) (see online version for colours)



2.3.2 *Feature extraction*

The stages to accomplish this explanation are raw data cleansing, feature selection, and feature extraction. Machine learning is concerned with extracting target-related relevant data from given selected features. Just those characteristics that contribute to the objective are appropriate in the machine learning algorithm when provided an attribute dataset and a target. Inconsequential features not only consume processing resources but also create huge noise.

To make machine learning more efficient and effective, we anticipate a relatively small feature component space, each contributing more to the prediction target. Feature extraction is a transformation that produces a new set of attributes.

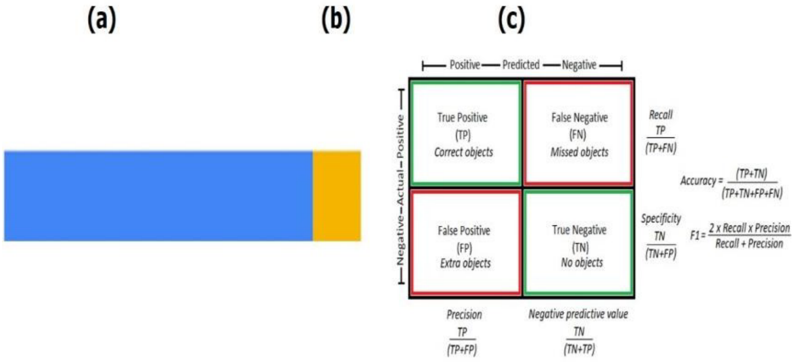
- Have a relatively small size.
- Have the most significant possible similarity with the target

Principal component analysis (PCA), independent component analysis (ICA), LDA, and Manifold are standard linear system algorithms. A wide range of Manifold-based algorithms, kernelled ICA, are used for nonlinear systems. Text and image sets of data frequently have large feature dimensions and highly correlated characteristics, which fit well with deep learning-based embedding or CNN, RNN-based algorithms. Most feature extraction methods researchers use in current works are PCA and LDA. They maintained almost 95% variance, which means only a 5% of loss of information.

2.3.3 Generate train and test set

The most basic way to evaluate the effectiveness of a machine learning system is to utilise separate training/testing data. Therefore, we may divide our initial data into two sections, as shown in Figure 11(a) and (b). We train the algorithms in Section 1, predict outcomes in Section 2, and then compare the forecasts to the expected results.

Figure 11 (a) Training set; (b) test set (Simon et al., 2016) and (c) confusion matrix (Sun and Huang, 2010) (see online version for colours)


(a)
(b)
(c)
Positive — Predicted — Negative
Negative – Actual – Positive
True Positive (TP) Correct objects
False Negative (FN) Missed objects
False Positive (FP) Extra objects
True Negative (TN) No objects
Recall TP/(TP+FN)
Accuracy = (TP+TN)/(TP+TN+FP+FN)
Specificity TN/(TN+FP)
F1 = 2 x Recall x Precision / Recall + Precision
Precision TP/(TP+FP)
Negative predictive value TN/(TN+TP)

2.3.4 Classification (ML/DL) models

The training dataset is fed into ML/DL algorithms. The input images transform into numerical (1D) data by building a feature learning convolving and pooling layer. Later changed data is given to the model to classify each input into respective output labels (Simon et al., 2016).

2.3.5 Evaluation

The performance metric of the model is based on a confusion matrix, which includes accuracy, precision, recall, AUC, and F1Score (Visa et al., 2011), as shown in Figure 11(c).

The steps of cancer detection of speech-related organs using ML/DL algorithms are explained based on the literature survey on cancer/disease detection of the Pharynx, larynx, voice disorders, and oral cancer. It helps to detect/diagnose at an early stage, which helps save human life and their voice/speech. Based on the literature studies and flow of cancer detection, we observe that cancer detection was held for each organ separately. Thus, this motivates us to develop a mobile application that detects the cancer of all speech-related organs. One more challenging aspect here is incorporating the images of the larynx, Pharynx, oral cavity, tongue, and lips to train the model. Inserting the camera to capture the images of speech organs feeds into the application, which shows the results as either cancerous or non-cancerous.

3 The proposed work

Each research in existing or current works has explored only one specific algorithm or feature selection on typical speech organ-related input image datasets. The image dataset used for the detection is noisy-free images. Thus, it got limited to a specific set of methodologies, as discussed in Chapter 3. Therefore, we are focused on different strategies and techniques that help produce the result with additional noisy image datasets.

To analyse and review speech-related cancer detection using algorithms because many billion people die from neglect detecting cancers or infections at early stages. Thus, this motivates us to explore more on this research area and build a mobile application that helps or usages for everyone to detect cancer of speech-related organs at one click of images using the mobile camera.

The dataset for the proposed work contains images of the Pharynx, larynx, and oral cavity, including lips and tongue, of all the age groups between 18 and 80 years old, with and without infection. The thermal images are captured with a FLIR thermal camera model SC620. Each infrared image has a dimension of 640×480 pixels. The images are collected from ENT hospitals with ~1Lakh samples with different resolutions. Our proposed work should give the best accuracy even if the input image is low resolution and noisy. The architecture diagram of the proposed work is mentioned in Figure 12.

Furthermore, it is essential to note that most of these appear to work but do not provide enough information about the database-separated technique during the training framework. Then there are two main approaches to consider. On the one hand, all patterns of each patient can be stacked in one database and split into train/test datasets. On the contrary hand, every patient's image sequence is assigned to either the train or test set, as shown in Figure 12. The methodology flow of the proposed work is shown in Figure 13.

The detailed workflow of fine-tuned and benchmark models is as follows. There are three phases:

1 describes the data acquisition, pre-processing, and data augmentation

2 shows the three core activities (baseline, benchmark, and fine-tuned CNN models)

3 displays the performance metrics used for evaluating all the CNN architectures as shown in Figure 13.

The exact size of the trained model will be freeze while fine-tuning the model based on accuracy for the respective dataset. But this proposed dataset contains all speech-related organs images with different resolutions and ages. The training dataset for the model is more than ~80,000 images are considered as input to train the model. The result will be either the input image is infected/non-cancerous or cancerous.

Figure 12 The architecture diagram of the proposed work (see online version for colours)

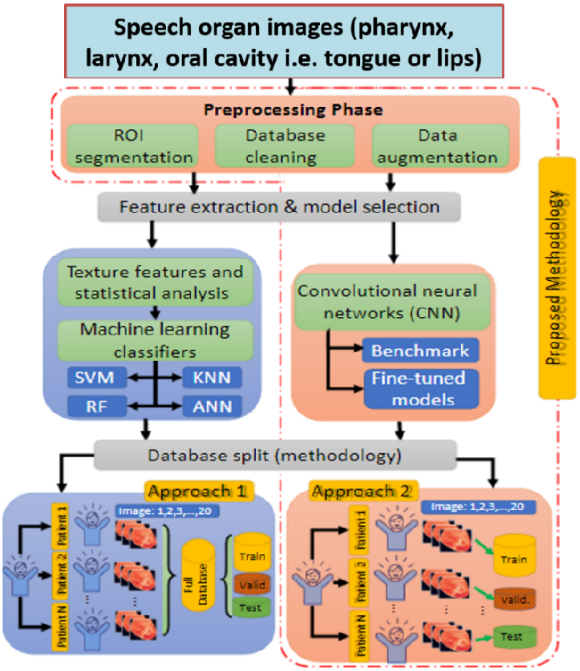
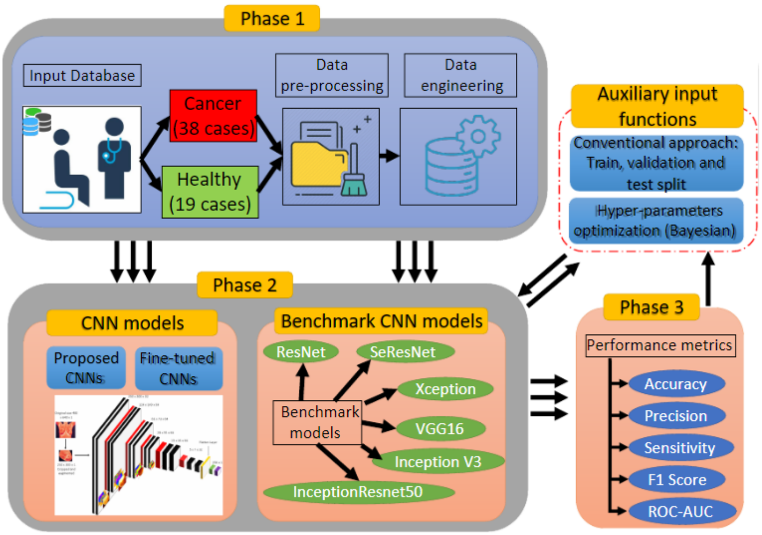


Figure 13 Methodology flow of proposed work (see online version for colours)



Thus, the future work has one more challenging aspect: incorporating the images of the larynx, Pharynx, oral cavity, tongue, and lips to train the model. Inserting the camera to capture the images of speech organs feeds into the application, which shows the results as either cancer or non-cancer.

4 Discussions

This work has addressed several larynges, pharyngeal cancer, voice problems, databases, feature extraction, and ML and DL algorithms in speech-related cancer detection research. However, the SVM classifier is the most extensively employed compared to other ML approaches. It might be because the SVM classifier performs better for high-dimensional and small-sized datasets, and getting massive databases for scientific purposes is challenging. On the other hand, ANN and CNN need massive data for training the models to enhance classification accuracy.

Most of the unique approaches or advances suggested in the research are primarily focused on disease characterisation methods, also known as feature extraction techniques. Many researchers claim to have proposed a novel approach for feature extraction and seen an outperformance. However, deep learning algorithms that extract the features intrinsically but need many training instances can bypass the feature extraction process. Even though deep neural networks (DNNs) has evolved as a robust ML approach, relatively few studies have been done utilising deep learning, which may be attributed to the lack of big datasets.

It has been discovered that most papers have shown high accuracy in cancer disease diagnosis, reaching 98% in a few situations. Nevertheless, the outcomes are database-specific. Therefore, it cannot be assured that the solution established for one database can be extended to other databases with the same efficiency.

The availability of big standard datasets is critical for research in this field and for developing real-time healthcare systems. To our understanding, only a few traditional benchmarking databases may be utilised for such research because such a database necessitates the involvement of doctors who are specialists in the field of voice problem diagnosis. The local database and automation process for a particular hospital may be created. Physicians generally use audio files for diagnosis and may not keep track of or preserve the audio files/images depending on the illness corrected in the long run.

5 Conclusion

Speech organ disease and cancer detection using ML/DL models have increased over the past few years. By this consideration, this paper has surveyed speech organ disease and cancer detection using advanced ML methods. This survey reviews the state-of-the-art approaches to speech organs and cancer detection based on ML methods. The processes such as pre-processing images, feature extraction, classification, and prediction in image-based speech organs and cancer diagnosis using several datasets are reviewed and discussed. The review compares state-of-the-art with the process, algorithm, features extracted, performance metrics and datasets used. The main goal of this survey is to encourage upcoming research in speech organ disease and cancer detection using ML-based methods at a very early stage to avoid the risk of death or loss of speech. In the future, the survey motivates us to develop early detection of cancer-related speech organs with meager resources of applications to use in real-time using mobile phones.

References

Abujamous, L., Tbakhi, A., Odeh, M., Alsmadi, O.A., Kharbat, F.F. and Abdel-Razeq, H. (2018) 'Towards digital cancer', *Genetic Counseling*, *1st International Conference on Cancer Care Informatics* (*CCI*) *2018*, pp.188–194.

Akshara, R. and Latchoumi, T.P. (2021) *Identification of Throat Cancer by Machine Learning*, Asurey.

Al-Maaitah, M. and AlZubi, A.A. (2018) 'Enhanced computational model for gravitational search optimized echo state neural networks based oral cancer detection', *Journal of Medical Systems*, Vol. 42, No. 11, p.205.

Askarian, B., Yoo, S.C. and Chong, J.W. (2019) 'Novel image processing method for detecting strep throat (streptococcal pharyngitis) using smartphone', *Sensors* (Basel, Switzerland), Vol. 19, No. 15, 3307, pp.1–17, https://doi.org/10.3390/s19153307

BenAicha, A. and Ezzine, K. (2016) 'Cancer larynx detection using glottal flow parameters and statistical tools', *ISIVC International Symposium Signal, Image, Video Commun.*, Vol. 2016, No. 1, pp.65–70, 2017.

Binti Mat Kasim, N.A., Binti Abd Rahman, N.H., Ibrahim, Z. and Abu Mangshor, N.N. (2018) 'Celebrity face recognition using deep learning', *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 12, No. 2, p.476, doi: 10.11591/ijeecs.v12.i2.pp476-481.

Bulbul, H. and Ünsal, Ö. (2010) 'Determination of vocational fields with machine learning algorithm', *Ninth International Conference on Machine Learning and Applications*, USA, pp.710–713.

Faden, H., Callanan, V., Pizzuto, M., Nagy, M., Wilby, M., Lamson, D., Wrotniak, B., Juretschko, S. and StGeorge, K. (2016) 'The ubiquity of asymptomatic respiratory viral infections in the tonsils and adenoids of children and their impact on airway obstruction', *International Journal of Pediatric Otorhinolaryngology*, Vol. 90, November, pp.128–132 (PMCFree article).

Fang, S.H., Tsao, Y., Hsiao, M.J., Chen, J.Y., Lai, Y.H., Lin, F.C. and Wang, C.T. (2019) 'Detection of pathological voice using cepstrum vectors: deep learning approach', *Journal of Voice*, Vol. 33, No. 5, pp.634–641.

Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C. and Parkin, D.M. (2010) 'Estimates of worldwide burden of cancer in 2008: GLOBOCAN2008', *International Journal of Cancer*, Vol. 127, No. 12, pp.2893–2917.

Follmann, D., Huang, C.Y. and Gabriel, E. (2016) 'Who really gets strep sore throat? confounding and effect modification of a time-varying exposure on recurrent events', *Statistics in Medicine*, Vol. 35, No. 24, 30 October, pp.4398–4412 (PMCFree article).

Fujimura, S., Kojima, T., Okanoue, Y., Shoji, K., Inoue, M., Omori, K. and Hori, R. (2022) 'Classification of voice disorders using a one-dimensional convolutional neural network', *Journal of Voice*, Vol. 36, No. 1, pp.15–20.

Georgopoulos, V.C. (2020) 'Advanced time-frequency analysis and machine learning for pathological voice detection', *Signal Process. CSNDSP. 12th International Symposium Communal System Networks Digit*, Portugal, pp.10–14.

Heo, J., Lim, J.H., Lee, H.R., Jang, J.Y., Shin, Y.S., Kim, D. and Kim, C.H. (2022) 'Deep learning model for tongue cancer diagnosis using endoscopic images', *Scientific Reports*, Vol. 12, No. 1, pp.1–10.

Inaba, A., Hori, K., Yoda, Y., Ikematsu, H., Takano, H., Matsuzaki, H., Watanabe, Y., Takeshita, N., Tomioka, T., Ishii, G., Fujii, S., Hayashi, R. and Yano, T. (2020) 'Artificial intelligence system for detecting superficial laryngo-pharyngeal cancer with high efficiency of deep learning', *Head and Neck*, Vol. 42, No. 9, pp.2581–2592.

Jeyaraj, P.R., Panigrahi, B.K. and Samuel Nadar, E.R. (2020) 'Classifier feature fusion using deep learning model for non-invasive detection of oral cancer from hyperspectral image', *IETE Journal of Research*, pp.1–12.

Jirawanitcharoen, K., Kiattisin, S., Leelasantitham, A. and Chaiprapa, P. (2009) 'A method of detecting to nsillitis images based on medical knowledge and neural network', *Proc 2nd IEEE International Conference Computability Science Inférieure Technologia ICCSIT2009*, pp.125–128.

Kim, G.H., Sung, E.S. and Nam, K.W. (2021) 'Automated laryngeal mass detection algorithm for home-based self-screening test based on convolutional neural network', *Bio Medical Engineering OnLine*, Vol. 20, No. 1, p.51.

Kim, H., Jeon, J., Han, Y.J., Joo, Y., Lee, J., Lee, S. and Im, S. (2020) 'Convolutional neural network classifies pathological voice change in laryngeal cancer with high accuracy', *Journal of Clinical Medicine*, Vol. 9, No. 11, 3415, pp.1–15.

Kono, M. (2020) 'Diagnosis of pharyngeal cancer on endoscopic video images by artificial intelligence', *United European Gastroenterology Journal*, Vol. 8, 8 SUPPL, p.163, https://www.embase.com/search/results?subaction=viewrecord & id=L634119899& from=export %0A, http://dx.doi.org/10.1177/2050640620927345

Mangileva, D., Dokuchaev, A., Khamzin, S., Zubarev, S., Lyubimtseva, T., Lebedev, D.S. and Solovyova, O. (2020) 'Removing artifacts from computed tomography images of heart using neural network with partial convolution layer', *Radio Electronics and Information Technology (USBEREIT) Ural Symposium on Biomedical Engineering*, Yekaterinburg, pp.195–198.

Mohammed, M.A., Abdulkareem, K.H., Mostafa, S.A., KhanapiAbd Ghani, M., Maashi, M.S., Garcia-Zapirain, B., Oleagordia, I., Alhakami, H. and AL-Dhief, F.T. (2020) 'Voice pathology detection and classification using convolutional neural network model', *Applied Sciences*, Vol. 10, No. 11, pp.1–13.

Nayak, G.S., Kamath, S., Pai, K.M., Sarkar, A., Ray, S., Kurien, J., D'Almeida, L., Krishnanand, B.R., Santhosh, C., Kartha, V.B. and Mahato, K.K. (2006) 'Principal component analysis and artificial neural network analysis of oral tissue fluorescence spectra: Classification of normal premalignant and Malignant pathological conditions', *Biopolymers*, Vol. 82, No. 2, pp.152–166.

Nocini, R., Molteni, G., Mattiuzzi, C. and Lippi, G. (2020) 'Updates on larynx cancer epidemiology', *Chinese Journal of Cancer Research*, Vol. 32, No. 1, pp.18–25.

Pahwa, K. and Agarwal, N. (2019) 'Stock market analysis using supervised machine learning', *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing* (*COMITCon*) *2019*, pp.197–200.

Parpulov, D., Samorodov, A., Makhov, D., Slavnova, E., Volchenko, N.N. and Iglovikov, V.I. (2018) 'Convolutional neural network application for cells segmentation in immune cytochemical study', *Radio Electronics and Information Technology (USBEREIT) Ural Symposium on Biomedical Engineering*, Yekaterinburg, pp.87–90.

Parveen and Singh, A. (2015) 'Detection of brain tumor in MRI images using combination of fuzzy c-means and SVM', *2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, India, pp.98–102.

Phensadsaeng, P. and Chamnongthai, K. (2017) 'The design and implementation of an automatic tonsillitis monitoring and detection system', *IEEE Access*, Vol. 5, pp.9139–9151.

Rabiner, L. and Juang, B.H. (1993) *Fundamentals of Speech Recognition*, Prentice Hall.

Sankaranarayanan, R., Masuyer, E., Swaminathan, R., Ferlay, J. and Whelan, S. (1998) 'Head and neck cancer: a global perspective on epidemiology and prognosis', *Anticancer Research*, Vol. 18, No. 6B, pp.4779–4786.

Sankaranarayanan, R., Ramadas, K., Thomas, G., Muwonge, R., Thara, S., Mathew, B., Rajan, B. and Trivandrum Oral Cancer Screening Study Group (2005) 'Effect of screening on oral cancer mortality in Kerala, India: a cluster-randomised controlled trial', *Lancet*, Vol. 365, No. 9475, pp.1927–1933.

Sharma, N. and Om, H. (2014) 'Using MLP and SVM for predicting survival rate of oral cancer patients', *Network Modeling Analysis in Health Informatics and Bioinformatics*, Vol. 3, No. 1, pp.1–10.

Sharma, N. and Om, H. (2015) 'Usage of probabilistic and general regression neural network for early detection and prevention of oral cancer', *The Scientific World Journal*, p.234191.

Simon, A., Deo, M., Selvam, V. and Babu, R. (2016) 'An overview of machine learning and its applications', *International Journal of Electrical Sciences and Engineering*, Vols. 22–24, pp.22–24.

Singh, V.P. and Maurya, A.K. (2021) 'Role of machine learning and texture features for the diagnosis of laryngeal cancer', *Mach. Learn. Healthc. Appl.*, pp.353–367.

Subramaniam, S., Jayanthi, K.B., Rajasekaran, C. and Sunder, T. (2019) 'Segmentation of RoI in medical images using CNN-A. comparative study', *TENCON*, *2019–2019 IEEE Region 10 Conference* (*TENCON*), New Zealand, pp.767–771.

Sun, S. and Huang, R. (2010) 'An adaptive k-nearest neighbor algorithm', *Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, China, Vol. 1, pp.91–94.

Tamashiro, A., Yoshio, T., Ishiyama, A., Tsuchida, T., Hijikata, K., Yoshimizu, S., Horiuchi, Y., Hirasawa, T., Seto, A., Sasaki, T., Fujisaki, J. and Tada, T. (2020) 'Artificial-intelligence-based detection of pharyngeal cancer using convolutional neural networks', *Digestive Endoscopy*, p.32, doi: 10.1111/den.13653.

Xiong, H., Lin, P., Yu, J.G., Ye, J., Xiao, L., Tao, Y., Jiang, Z., Lin, W., Liu, M., Xu, J., Hu, W., Lu, Y., Liu, H., Li, Y., Zheng, Y. and Yang, H. (2019) 'Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images', *EBiomedicine*, Vol. 48, pp.92–99.

Yoo, T.K., Choi, J.Y., Jang, Y., Oh, E. and Ryu, I.H. (2020) 'Toward automated severe pharyngitis detection with smartphone camera using deep learning networks', *Computers in Biology and Medicine*, Vol. 125, August, p.103980.

Website

Statistics on Voice Speech, and Language|NIDCD.nih.gov