Sumer Journal for Pure Science (ISSN: 2790-7031)
SJPS 2023 (December), Vol. 2, No. 2, p.p.: 88-102

COVID-19Disease Diagnosis using Artificial Intelligence based on Gene Expression: A Review

Qusay K. Kadhim[1]*, Sanaa H. Dhahi[2], Ekhlas G. Abdulkadhim[3], Wasan A.Wahab Alsiadi[4]

[1]Department of Computer Science, College of Science, University of Diyala, Diyala, Iraq
[2]Department of Hotel Studies, Collage of Tourism Sciences, University of Kerbala, Kerbala, Iraq
[3]Department of Tourism Studies, Collage of Tourism Sciences, University of Kerbala, Kerbala, Iraq
[4]Department of Biology, College of Education for Pure Science, Ibin Al-Haitham, University of Baghdad, Baghdad, Iraq

Abstract

Clinical symptoms of COVID-19, which can cause pulmonary inflammation, are varied. The use of microarray technology to examine variations in gene expression in particular organisms has become a popular new trend in genetic research, which can be used for disease detection and prediction. Large numbers of unprocessed raw gene expression profiles can occasionally make it difficult to choose dataset attributes and categorize them into the proper group or class, which can lead to computational and analytical challenges. Using the whole set of genes makes getting acceptable COVID-19 classification accuracy difficult due to the oversized dimensions, noise in the gene expression data and tiny sample sizes. This review thoroughly assesses microarray COVID-19 illness investigations, emphasizing techniques for selecting features. Hence, COVID-19 could be controlled with the development of a very sensitive and accurate point-of-care COVID-19 detection device, and this review thus offers general recommendations for researchers and workers in the biosensor field.

Keywords: COVID-19 Disease, Feature Selection, Microarray Technology, Machine Learning, Deep Learning, Gene Expression, SARS-CoV-2.

تشخيص مرض كوفيد-19 باستخدام الذكاء الاصطناعي بناءً على التعبير الجيني: مراجعة

قصي كنعان كاظم[1]*, سناء حماد ضاحي[2], اخلاص غالب عبدالكاظم[3], وسن عبد الوهاب السعيدي[4]

[1]قسم علوم الحاسوب, كلية العلوم, جامعة ديالى, ديالى, العراق

[2] قسم الدراسات الفندقية ,كلية العلوم السياحية ,جامعة كربلاء ,كربلاء ,العراق

[3]قسم الدراسات السياحية, كلية العلوم السياحية ,جامعة كربلاء ,كربلاء ,العراق

[4] قسم علوم الحياة , كلية التربية للعلوم الصرفة (ابن الهيثم)، جامعة بغداد، بغداد، العراق

الخلاصة

تتنوع الأعراض السريرية لكوفيد-19، والتي يمكن أن تسبب التهابات تنفسيه. أصبح استخدام تقنية المصفوفات الدقيقة لفحص الاختلافات في التعبير الجيني في كائنات معينة اتجاهًا جديدًا شائعًا في الأبحاث الجينية.،والذي يمكن استخدمه في الكشف عن الأمراض والتنبؤ بها. قد تؤدي الأعداد الكبيرة من ملفات تعريف التعبير الجيني الخام غير المعالجة في بعض الأحيان إلى صعوبة اختيار سمات مجموعة البيانات وتصنيفها في المجموعة أو الفئة المناسبة، مما قد يؤدي إلى تحديات حسابية وتحليلية. إن استخدام مجموعة الجينات الكاملة يجعل الحصول على دقة تصنيف مقبولة لكوفيد-19 أمرًا صعبًا بسبب الأبعاد الكبيرة والتشويش في بيانات التعبير الجيني وأحجام العينات الصغيرة. تُقيّم هذه المراجعة بدقة التحري عن مرض كوفيد-19 باستخدام المصفوفة الدقيقة، مع التركيز على تقنيات اختيار الميزات. بالتالي يمكن السيطره على كوفيد-19 من خلال تطوير جهاز كشف كوفيد-19 حساس ودقيق للغاية في النقطة المطلوبة، وبالتالي تقدم هذه المراجعة توصيات عامة للباحثين والعاملين في مجال أجهزة الاستشعار الحيوية .

* Dr.qusay.kanaan@uodiyala.edu.iq

1. Introduction

There are now major concerns to public health all over the world due to the recent appearance of the new coronavirus (SARS-CoV-2, 2019- nCoV), which is what triggered the coronavirus disease 2019 (COVID-19) outbreak in China [1]. Although travel to and from Wuhan and a number of other Chinese cities was prohibited starting on January 23, 2020, in an effort to stop the spread of the virus both within China and globally, as of July 4, 2020, there were 11,191,872 cases of laboratory-confirmed SARS-CoV-2 infection and 529,122 deaths reported globally [2].The lack of efficient point-of-care testing (POCT) assays to quickly and accurately detect SARS-CoV-2-infected individuals may be contributing to the situation's escalating seriousness. Additionally, SARS-CoV-2 infected patients who are asymptomatic or pre-asymptomatic are highly contagious, and due to the lack of reliable detection tests, many of these patients have come into touch with uninfected individuals before being quarantined at home or admitted to a hospital [3]. Additionally, the SARS-CoV-2 outbreak and the flu season happened at the same time. Due to the high rate of nosocomial transmission of the virus, the concurrent visits of influenza patients to hospitals also contributed to the growing spread of the SARS-CoV-2 infection[4]. Therefore, a quick, affordable, and highly accurate POCT approach is critical for prompt isolation of infected individuals and efficient contact tracing of possible SARS-CoV-2 infected cases[5].

We first describe the SARS-CoV-2 genome, gene expression features, and viral particle structure in this review. Afterward, a summary of the available SARS-CoV-2 RNA, viral particle, antigen, and antibody detection techniques is shown Figure 1. Speaking of unsolved issues, the issues with false-positive and false-negative outcomes in clinical practice are also covered [6]. The paper also describes innovative nanoparticle-based lateral flow assay, electrochemical biosensors, and microfluidic chips as potentially viable novel detection methodologies that may be used in the future to enhance the efficacy of COVID-19 detection assays. Finally, we address future research ideas for utilizing portable detection technologies to build point-of-care SARS-CoV-2 detection systems that are highly accurate, affordable, and simple to use [7].

The application of machine learning approaches for COVID-19 Disease diagnosis has considerably improved over the past ten years[8]. Deep Learning (DL) and support vector machines are the two classification techniques that are most frequently utilized. The major distinction between SVM and ANN is the nature of an optimization problem [9]. The SVM offers a globally ideal answer. In contrast, ANN offers a locally optimal solution a crucial step in both SVM and ANN is feature extraction [10]. However, the learning model directly incorporates the feature extraction process in deep learning [11] . When working with large datasets, for image data, deep learning has been shown to be extr emely useful. In order to improve the accuracy of COVID-19 illness classification, several researchers used ensemble methods[12].

Reduced amounts of the Microarray dataset's duplicated and useless data is used in dimensionality reduction (DM), which seeks to improve the accuracy of a classification algorithm reducing a system's dimensionality, can be done in a variety of ways [13]. The application domain and unique characteristics of the dataset influence the dimensionality reduction techniques. Feature selection strategies come in the form of filter, wrapper, embedding, and hybrid approaches [14]. In accordance with the traits of specific users, filter algorithms choose qualities [15]. With the use of population changes or machine learning, a subset of features is chosen using wrapper techniques.

The capacity to conduct calculations rapidly at the expense of accuracy is well-known for methods that use filters, while wrapper techniques need less work and have higher accuracy and performance. Filter-based strategies outperformed wrapper methods in fields with huge datasets[16]. Both methods have the flaw of not taking the relationship between the classifier and addiction among them depending on which features are employed, the various features produce differing categorization accuracy [17]. Embedded methods, available Algorithms for learning can help you perform better. Wrapper approaches need more processing, but embedding techniques have the advantage of interacting with the classification system [18].

The purpose of this work is to discuss the difficulties and issues surrounding microarray datasets for COVID-19 disease, as well as the current feature selection techniques used in feature selection. Give a thorough explanation of the microarray experiment and list the shortcomings of the available techniques. This publication also offers significant future directions for this area of research. The review paper contains all the following details: Section 2 gives a general description of the Microarray technology and related data, and Section 3 is dedicated to Gene Selection. Reduced Dimension in Section 4, approaches are categorized based on their taxonomy. Opened research issues are explored in Section 5. When gathering the data, these worries were taken into account. Section 6 covers the literature review, and sections 7 and 8 contain the paper's discussion and conclusion.

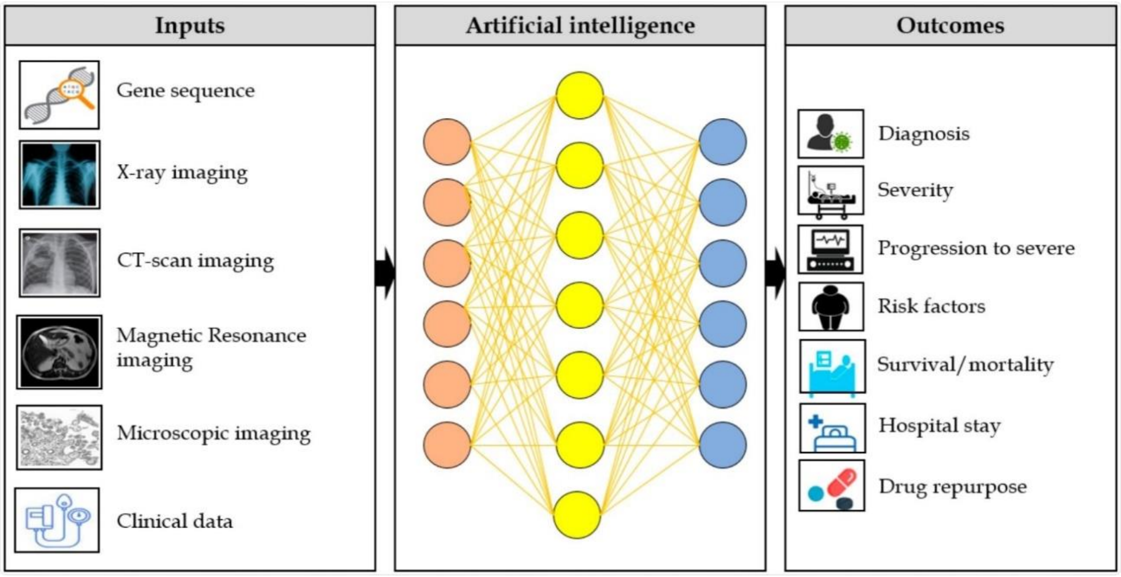

Figure -1 Artificial intelligence Techniques to fight Covid-19 [19]

2. Review of SARS-CoV-2

The initial step in managing COVID-19 is the quick and accurate detection of SARS-CoV-2 made feasible by real-time reverse transcription-polymerase chain reaction (RT-PCR). RT-PCR can identify nasopharyngeal fluids having SARS-CoV-2 nucleic acids in them [20]. Testing is used to stop the spread of infectious diseases within communities that contain asymptomatic sick people, whose viral shedding may unintentionally spread the infection to elderly people and others with preexisting medical conditions. For the COVID-19 pandemic to be contained, accurate viral detection is a good place to start. False-negative test results help illness spread, which compromises public safety. Serological testing complements virus identification, suggesting prior infection, which could be leveraged for therapeutic advantage. Improving test sensitivity and specificity remains a critical need [21]. Through the employing qualitative detection antibodies, antibodies are found using an enzyme-linked immunosorbent assay. These tests identify an immunological reaction to the viral spike protein and may be used to evaluate protection against further viral exposure and/or for contact tracing [22]. The significance of such testing can therefore not be emphasized. This holds true for epidemiological analyses and the vast majority of the world's unmet medical needs14. In the future, diagnostic tests to boost immunoassay sensitivity and specificity will be developed. As reinfections appear, such testing will, in fact, eventually indicate viral protection. The next step in controlling COVID-19 is to develop immunity against SARS-CoV-2. To this aim, our goal in this Review is to provide an overview of the clinical disease presentation with an emphasis on the most effective use of diagnostic tests based on nanomaterials at the individual, communal, and societal levels. The Review describes existing and upcoming COVID-19 nanomaterial diagnostics. The goal is to make it easier to stop the virus's global spread [23].

3. Microarray Technique

Since its inception, Microarray Technology (MT) has made significant strides within the field of biology. It's regarded as a platform to significant novel investigations. It now has the ability to simultaneously analyze over a hundred thousand genes activity. Even, if the majority of scientists and biologists struggle to mine and deal with this type of data. The outcomes of Microarray investigations are accessible through a number of databases[24]. Microarrays have been used in scientific study since the middle of the 1980s. Augenlicht et al. (1987) originally described DNA microarrays after finding over 4000 complementary DNA (cDNA) sequences on nitrocellulose. Tens of thousands of genomes may be examined and their expression measured simultaneously by biologists thanks to microarrays. The development of the technique has benefited microarray research, bioinformatics, and other medical domains. This kind of microarray, which has numerous tiny genetic material attached to a solid surface in small regions, is frequently denoted to as a "biochip" or "DNA chip". DNA Scientists use microarrays as a platform for examining the locations where several genes are expressed at once, diverse elements that make up an individual's genotype [25].

3.1 Microarray data

Microarray experiment data are often organized and saved in large matrices (XY)[26]. Each Microarray data matrix (feature) contains samples in columns and genes in rows, as shown in Table 1. The matrices X by Y that hold the microarray data are enormous. A sample's cells each have a unique value expressing a particular gene, where Y is the number of columns and X is the number of rows. Assuming that i an integer between j and X and that j is a negative number between 1 and X, L ij displays the gene expression levels in relation to the case study or taster j.

Table 1- Microarray data matrix

Samples	Gene Expression		
	G1	***G2***	***GM***
S1	Y11	Y12	Y1M
S2	Y21	Y22	Y2M
-			
-			
SN	YN1	YN2	YNM

4. History and Development of Gene Selection

Gene selection is a method for removing redundant and/or useless genes from gene expression data; one illustration of this is the DNA microarray[27]. Machine learning is used in feature selection to choose gene; it is perfect for applications requiring many traits. In order to uncover and express the most helpful information, researchers first employ Gene Selection techniques to find and delete identical genes in the original location. Increasing the number of genes, over-fitting should theoretically result in decreased generalization and decreased model performance. Work we're doing now for Gene Selection (GS) focuses more on determining which genes are crucial than it does on trimming unnecessary or redundant genome. To produce significant results, relevance, redundancy, and complementarity must be given top priority. Whether or whether a gene has the necessary knowledge about the specified class determines its significance. The feature set can be categorized into three groups, according to: highly relevant, somewhat relevant, and irrelevant. Characteristics that are redundant and those that aren't are the two different kinds of marginally relevant characteristics[28]. The majority of the pertinent information is found in the non-redundant and highly important feature sections. Similar algorithms based on microarray data are used to select genes, as shown in Figure 2.

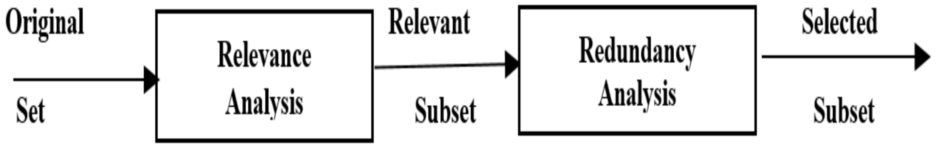

Figure -2 Gene Selection Framework[26].

5. Dimensional Reduction Strategies

Classification algorithms encounter various computational and memory difficulties when working with significant volumes of dimensional data [29]. Dimensionality reduction (DR) and feature transformation (FT), often known as extraction and selection of attributes are two ways to make a system's dimensions smaller. When employing the feature selection approach, there is no indication of the significance of a single feature absent; only when numerous distinctive traits are required does this not apply, in which case the selection of some characteristics could result in the loss of information group. In contrast, using feature extraction, one can reduce the feature set without substantially dropping the original feature's information. The choice and feature selection of feature extraction methodologies depends on the data type and application domain.

5.1 Feature selection

Big dimensional datasets contain redundant, deceptive, or both features, which complicate further interpretation of the data and do not advance the learning process[30]. When you select the most important features out of all those that can be used to distinguish between classes, you use a feature subset selection process. A particular notion of significance activates the feature selection algorithm, a statistical technique. Numerous feature selection techniques have been empirically tested [31]. The term "search problem" is often used to refer to feature selection according to different evaluation criteria. Exponential, sequential, or random search approaches characterize the search organization used by feature selection algorithms. To construct successors, one can experiment with five different operators; weighted, compound, among them are only a couple at random and. Assessment Metrics: Analysing the stability, uncertainty, and information in Figure 2 illustrates how successors can be evaluated using probability mistake, divergence, dependence, and the distance between classes. The three basic categories of feature selection strategies are filters, wrappers, and embedded/hybrid approaches. Due to a fact that a Feature Selection procedure is particular for the classifying being utilized, wrapper-based approaches enhance techniques that use filters. However, using Wrapper approaches in large feature spaces is very expensive due of their expensive processing fees and the requirement that a trained classifying be employed the selection of features requires extra time since each feature set must be evaluated. Though their classification reliability is poor, filtering techniques are more effective and quicker to calculate than conventional approaches, making them more suited for large, complicated datasets than wrapper approaches[32]. Recently, techniques for merging hybrid and embedded, which incorporate what filters and wrappers do best, have been developed. A hybrid technique for a feature subset combines independent tests with performance assessment tools[33]. There are two categories of filtering methods: Figure 3 shows feature weighting strategies and subset selection procedures in particular. Methods for weighing in take into account each element individually and give it a value based on how significant it is to the overall goals.

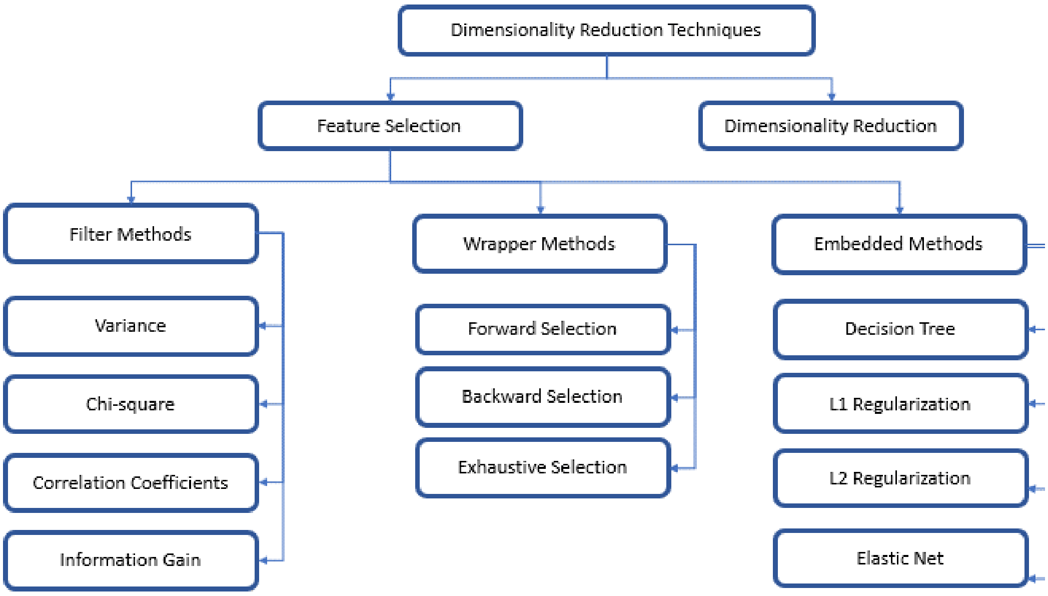

Figure -3 Approaches to dimension reduction [34]

5.2 Extracting Feature Transforming

The process of extracting features comprises transforming the initial attributes into more important ones. The following definitions fit the situation, extracted features the technique of combining continuous characteristics into linear combinations with strong group discrimination is known as "feature extraction." A prominent area of research neural networks and artificial intelligence is finding an appropriate representation for multivariate data. The data can be made simpler in this situation by using features extraction, which offers a linear combination of all the feature set's variables using input of the original input variable. A non-parametric methodology can be used to find the most important facts in a difficult and duplicated data collecting, making principal component analysis an easy-to-understand study [35]. We can increase information (measured by variance) and decrease duplication (measured by covariance) in our data by using Principal Component Analysis (PCA). Through a number of PCA techniques, information can be gained, wrapper approaches taken, or features extracted (e.g., from email data or drug detection data) are just a few of the alternate methods of dimension reduction that have actually been developed, and its effectiveness on two different types of data has been examined the outcomes of PCA feature extraction (transformation) are strongly correlated with the type of data used. When compared to information gain, the process used to choose which features to include in Wrapper for both types of data has a moderate effect on categorization accuracy. A study has demonstrated the importance of dimensionality reduction. Wrappers provide smaller feature subsets with improved classification accuracy when compared to feature extraction methods and feature selection methods. Wrappers are a good alternative to feature extraction techniques, albeit being computationally more expensive. In order to improve classification effectiveness, presented approaches for reducing the bi-level feature selection and extraction process' dimension. The first stage in dimension reduction is to select features based on how closely they connect to one another. In order to extract more features at the second level of PCA and LPP, certain first stage features are used. The effectiveness of the proposed strategy was evaluated on a wide range of frequently used datasets. According to the results, the suggested system performs better than single-level dimensionality reduction methods [36]. Figure 4 displays the categorized Artificial Intelligence (AI) techniques machine learning and deep learning selected for this review paper.

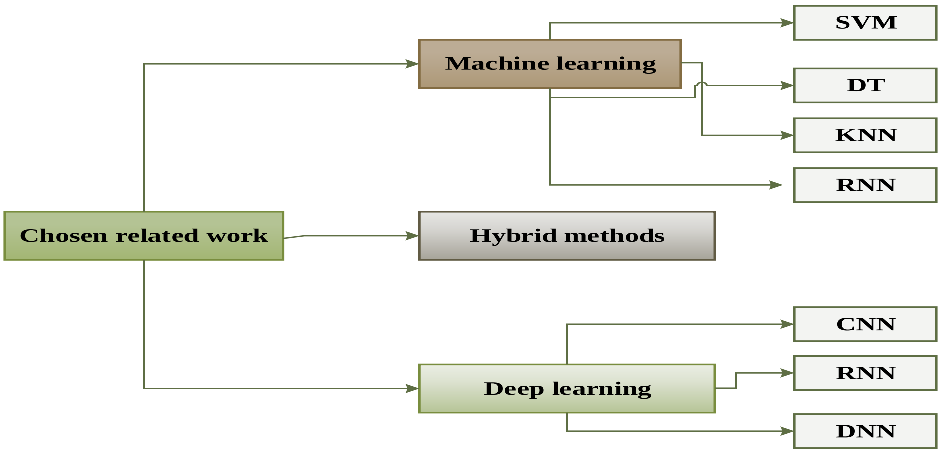

Figure -4 Classifying related works.

6. Literature Review

Several methods for treating COVID-19 disease are used in a variety of works in the literature. Examples from the most recent studies in the subject will be included in this section.

Alazab et al., (2020) [37], An artificial intelligence method that used two real-world datasets obtained from Australia and Jordan, a deep CNN was built to identify patients with COVID-19. One hundred X-ray images of actual patients were used to test their procedure. The approach has a 94%–98.83% accuracy rate for identifying COVID-19 patients in studies. The number of COVID-19 patients, cases of recovery, and fatalities over the course of the next seven days were also predicted using their method employing two predicting methodologies; specifically, the LSTM and the autoregressive integrated moving average model. Australia and Jordanian data sets were utilized for testing and training.

Apostolopoulos et al., (2020)[38], The usefulness of extracted features in categorizing X-ray indicators for the COVID-19 virus using the CNN named Mobile Net was tested. With the help of deep CNN, picture for COVID-19, biomarkers may be automatically extracted from X-ray images. The findings showed that COVID-19 could be classified into seven separate categories with an accuracy of 89.66%, and that COVID-19 and non-COVID-19 could be distinguished with an accuracy of 96.18%, 96.36%, and 98.42%, respectively.

Cheng et al., (2020)[39] , Provinces were inspected in the past. First, the Deep Learning (DL) approach was pertained using 4106 patient CT data. The images proved useful for analyzing lung features. According to the length of qualifying DL framework, 1266 patients (924 with COVID-19; 471 had follow-up of more than 5 days), and externally validated success of DL framework) and 342 patients with pneumonia were enrolled from six towns or regions. In the four prior sets of validation, the DL technique successfully distinguished COVID-19 cases from other pneumonia (AUC 0.87 and 0.88) and COVID-19 cases (AUC 0.86). When the patients were categorized using the DL technique, the length of time spent in the hospital varied significantly, limiting several characteristics.

Bansal,et al., (2020)[40], The onset of a serious sickness is associated with death from substantial alveolar injury and progressive respiratory failure. The absence of RT-PCR testing services during a pandemic crisis may also postpone the subsequent clinical judgment and treatment. For COVID-19 patients, chest imaging using CT is a useful testing and prognostication technique. This work proposed a poorly regulated deep learning system with CT scans to detect and classify COVID-19 infections. The technique may be helpful for eliminating manual CT scan markings and differentiating COVID-19 cases from non-COVID-19 cases. For the purpose of adjusting to changes in the authors created a multiple-scale learning system based on the size and location of lesions. The Decision Tree (DT) was fed function charts and classification layers. Results from radiologists and those from the suggested technique were compared (97. 82%).The training set's potential for high levels of noise, which could have an effect on model performance, is the study's shortcoming.

Babukarthik, et al., (2020)[41], XGBoost classifier was used to develop a prediction model for COVID-19 patient ventilation demands. A number of decision trees are used by the XGBoost classifier, and the results are combined into a single score. For this investigation, data on patients who registered with On 24 March 2020 and 4 May 2020, patients were either admitted to these five hospitals were gathered from five US healthcare systems. Twelve different factors, including temperature, blood urea nitrogen, and diastolic blood pressure, are monitored for every patient. When compared to the Modified Early Warning Score (MEWS), the suggested approach produces diagnostic ratios for ventilation prediction that are higher. Along with this accomplishment, the suggested model also yielded results with greater accuracy (90.67%).

Antonio et al., (2020)[22], the presence of symptoms and indicators was assessed for Jordan's normal and COVID-19 patients in both groups using data from the questionnaire. The researchers created a COVID-19 dataset that included the indications and symptoms of numerous patients. In order to predict future COVID-19 patients, the researchers used this dataset as input to a range of Machine Learning (ML) models (SVM, Multi-Layer Perceptron [MLP]). The most effective result in terms of classification accuracy was made by MLP (91.62%). SVM had the highest precision performance (91.67%). A combination of feature selection strategies and local search techniques was utilized to increase accuracy due to sample size restrictions and low precision.

Therapy, et al. (2020),[42], Applied ANN, KNN, and SVM classifiers to the Leukemia dataset after using the Bayesian feature selection strategy, the accuracy rate was 94.99%. Cho et al [9] used Uncorrelated Linear Discriminant Analysis (ULDA) to choose the features, and collected more accurate classification data (76.3%) for SVM, and Recursive Feature Elimination (RFE).

Zeng, et al., (2020)[34], The study's findings are related to the application of Internet of Technology (IoT) in healthcare settings, aiding in accurate medical diagnosis and emphasizing the standard of care provided to patients. Additionally, relying on IoT solutions for remote diagnosis reduces irregular hospital patient evaluations. An application to healthcare organizations would also produce accurate information on the illnesses that patient’s experience, involving them in the development of clinical research to produce more in-depth results. The Internet-based healthcare nursing classification was offered in this study. Researchers have suggested K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), and Neural Network (NN) algorithms as machine learning methodologies.

Patterson ,et al., (2021)[43], Completed a study that classifies COVID-19 against other coronaviruses using information gleaned from genomic sequences and a variety of machine learning-based classification techniques. SVM, naive Bayes, K-nearest neighbor, Random Forest, and Decision Tree are some of these classification techniques. Employ the Novel Coronavirus Resource Database for 2019. According to the results, the Decision Tree has a classification accuracy of 93%.

Kavitha, et al., (2021)[44], It was suggested that the Deep Learning Convolutional Neural Network (DCNN) model transform the mathematical properties using a DCNN technique, fill them into the picture domain from the new dataset. According to the trial findings, the suggested DCNN method for management categorization achieved 98.65% accuracy, compared to 98.65% for the traditional machine learning (Ensemble) algorithm. The results of the investigation into the type of therapy and the degree of treatment attentiveness prediction are confirmed by the performance measures.

Hasan et al., (2022)[45], the use of deep learning techniques to find COVID -19 biomarkers is the main focus of this study. Datasets from the ADNI-1 imaging study were utilized to generate genetic information on the COVID-19 disease using artificial neural networks (ANN) and convolutional neural networks (CNN) methods. In the whole genome approach of ADNI-1, the ANN and CNN algorithms achieved overall accuracy of 91% and 87%, respectively. The results indicate that classification algorithms are useful for early COVID-19 disease detection; however, selecting the best features within a certain price range can take a lot of effort.

Alimadadi, et al. ,2023[46], To anticipate the inference of significant cancer genes, they proposed a technique for creating machine learning models. Several Convolutional Neural Network (CNN) models were presented that classify tumor and non-tumor samples into particular or typical forms of cancer using uncontrolled gene expression inputs. Three CNN models were implemented: 1D-CNN, 2D-Vanilla-CNN, and 2D-Hybrid-CNN. These models were based on various designs of gene insertion and convolutional plots. The gene expression patterns from 10,340 samples from 33 different cancer types and 713 normal tissues that were matched with the Cancer Genome Atlas (TCGA) were used to train and evaluate the models. These models have a 95% forecast success rate.

Table 2 focuses on the feature selection strategy and summarizes the most recent improvements in the COVID-19 prediction system to offer us a quick overview of the work that has been done in this significant medical arena.

Table 2 list of the literature

References	Dataset	Gene Selection methods	Methods	Accuracy
Alazab et al., (2020) [35]	GSE5281	LSTM	LSTM	98.83%
Apostolopoulos et al., (2020) [36]	Not available to the general public.	CNN	CNN	96.18%
Cheng et al., (2020) [37]	ANM1, ANM2 GEO:GSE60862)	DL	DL	88%
Bansal,et al., (2020) [38]	Not public	DT	DT	97.82%
Babukarthik, et al., (2020) [39]	GSE1295	XGBoost	XGBoost	90.67%

Antonio et al., (2020) [20]	GSE1297	Bayesian	SVM, Multi-Layer Perceptron [MLP]	91.67%
Therapy, et al., (2020) [40]	GSE1297 GSE4757	ULDA	ANN, KNN and SVM	94.99%
Zeng,et al., (2020) [32]	GDS4602 and GDS460	SVM	SVM-RFE	76.3%
Patterson, et al., (2021) [41]	GSE63600& GSE63500	Ensemble	SVM, Naive Bayes, KNN, RF, and DT	93 %
Kavitha, et al ., (2021) [42]	Not available	DCNN	DCNN	98.65%
M Hasan et al., (2022)	ADNI-1	ANN & CNN	ANN & CNN	91%
Alimadadi, et al. ,(2023) [43]	ADNI-1	CNN	CNN	95%

7. Discussion

As time passes, more studies using diverse approaches have been published to forecast the onset of COVID-19disease. As a result of its significance, thorough evaluations of the state of research and application are needed. Therefore, this study's objective is to provide an in-depth summary of the newest research on COVID-19 illness detection using a variety of methodologies. A lot of study has been done on COVID-19disease from the end of 2011 until the present. According to a review of the relevant work, Approaches to feature extraction were shown to be much more suitable for automatically because of the noisy data for identifying COVID-19 disease than feature selection methods. Considering that noisy data is preferable to useless or duplicated data makes up the majority of biological datasets. In a number of applications, the tool of feature selection can be used to eliminate characteristics that are unnecessary or irrelevant. No particular method exists for choosing features that can be applied to all applications. There are some methods for removing unnecessary traits while avoiding duplicating features. A feature weighting method that only considers relevance falls short in addressing the demand for feature selection. Feature subset search techniques use an evaluation metric to determine which candidate feature subsets are the best.

Both irrelevant and redundant traits can be successfully eliminated using the consistency measure and two well-liked evaluation tools are the connection measure. The number of iterations required to identify the optimum feature subset is typically at least quadratic to the amount of features, according to experiments. Because of this, current subset search techniques with quadratic or more dimensional time complexity are not sufficiently scalable to handle large amounts of data. Two different kinds of feature selection strategies are filters and wrappers. Due to the fact that the feature selection procedure is adapted to the selected classification methodology, wrapper approaches often outperform filter methods. But because each feature set needs to be assessed using the trained model, many features are often too expensive to utilize. Large data sets are more suited for filter approaches than wrapper methods since they are so much faster. Recently, solutions using a hybrid paradigm have been put out to handle high-dimensional data by combining the advantages of both paradigms. Additionally, there are just a few methods for handling noisy data. To lessen the effect of type noise on the learning process, feature extraction methods have been suggested as a preprocessing stage. The type of data has a significant impact on the classification accuracy that can be achieved with different feature reduction techniques, according to studies. Methods for dealing with redundant and unnecessary features simultaneously are much more resilient and advantageous for the learning process than approaches that discretely handle redundant and/or irrelevant features. Therefore, research based on scant data would not be considered a substantial contribution to this field. There are three main drawbacks to diagnosing COVID-19 disease using different methods. The first is a data difference, which can be corrected in upcoming work via enhancing the model with extra features or knowledge-based traits. The second challenge was handling many data; cloud computing is used in this case would be more advantageous than large amounts of data locally be trained, requiring additional technical and labor-intensive tasks. A fourth obstacle, which is currently the major concern in this field, is the lack of readily available datasets.

8. Conclusion

A widespread use from machine learning algorithms is identifying COVID-19 disease. DNA microarray data present several challenges to machine learning research because of their extensive dimensional features and small sample numbers. The only other benefits of using feature selection as a pre-processing technique are reducing quantity of input features and memory and processing time savings. The accurateness of classification is increased with feature selection. The data's asymmetrical distribution of categories is another issue that researchers must address. There have been many test and training datasets identified, but in addition to the issue, the use of excessive features for a number of small samples, and the existence of outlier (also known as dataset shift) continues to be of concern. Researchers create numerous new techniques every year to improve earlier approaches' classification accuracy and overcome their shortcomings. Researchers also seek to help biologists identify and comprehend the primary mechanism that links genetic expression and illness.

Feature selection is used to overcome this issue, and the results are encouraging. Hybrid feature selection approaches are becoming more and more popular among researchers as a tool for feature selection assignments. These methods fundamentally fall under filters, wrappers, and embedding strategies. Filtering techniques are the most frequent since large datasets require a lot of computing power. The use of wrapper and embedding technologies has been carefully avoided.

References

[1] S. Karakanis and G. Leontidis, “Lightweight deep learning models for detecting COVID-19 from chest X-ray images,” *Comput. Biol. Med.*, vol. 130, no. September 2020, p. 104181, 2021, doi: 10.1016/j.compbiomed.2020.104181.

[2] X. Wu, C. Chen, M. Zhong, J. Wang, and J. Shi, “COVID-AL: The diagnosis of COVID-19 with deep active learning,” *Med. Image Anal.*, vol. 68, p. 101913, 2021, doi: 10.1016/j.media.2020.101913.

[3] S. Ouf and N. Hamza, “The Role of Machine Learning to Fight COVID-19,” *Int. J. Intell. Eng. Syst.*, vol. 14, no. 2, pp. 121–135, 2021, doi: 10.22266/ijies2021.0430.11.

[4] E. Luz *et al.*, “Towards an effective and efficient deep learning model for COVID-19 patterns detection in X-ray images,” *Res. Biomed. Eng.*, 2021, doi: 10.1007/s42600-021-00151-6.

[5] V. Mergen *et al.*, “Deep learning for automatic quantification of lung abnormalities in COVID-19 patients: First experience and correlation with clinical parameters,” *Eur. J. Radiol. Open*, vol. 7, 2020, doi: 10.1016/j.ejro.2020.100272.

[6] M. M. Rahaman *et al.*, “Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches,” *J. Xray. Sci. Technol.*, vol. 28, no. 5, pp. 821–839, 2020, doi: 10.3233/XST-200715.

[7] H. Yasar and M. Ceylan, “A new deep learning pipeline to detect Covid-19 on chest X-ray images using local binary pattern, dual tree complex wavelet transform and convolutional neural networks,” *Appl. Intell.*, vol. 51, no. 5, pp. 2740–2763, 2021, doi: 10.1007/s10489-020-02019-1.

[8] S. Annunziata *et al.*, “Impact of the COVID-19 pandemic in nuclear medicine departments: preliminary report of the first international survey,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 47, no. 9, pp. 2090–2099, 2020, doi: 10.1007/s00259-020-04874-z.

[9] H. A. A. Mohammed, I. Nazeeh, W. C. Alisawi, Q. K. Kadhim, and S. T. Ahmed, “Anomaly Detection in Human Disease: A Hybrid Approach Using GWO-SVM for Gene Selection,” *Rev. d’Intelligence Artif.*, vol. 37, no. 4, pp. 913–919, 2023, doi: 10.18280/ria.370411.

[10] M. Manav, M. Goyal, and A. Kumar, “Role of optimal features selection with machine learning algorithms for chest X-ray image analysis,” *J. Med. Phys.*, vol. 48, no. 2, pp. 195–203, 2023, doi: 10.4103/jmp.jmp_104_22.

[11] S. T. Ahmed and S. M. Kadhem, “Predicting Alzheimer’s Disease Using Filter Feature Selection Method,” *Iraqi J. Comput. Commun. Control Syst. Eng.*, vol. 22, no. 4, pp. 13–27, 2022, doi: 10.33103/uot.ijccce.22.4.2.

[12] M. Tabaa, H. Fahmani, M. El Ouakifi, and H. Bensag, “Covid-19’s rapid diagnosis open platform based on X-ray imaging and deep learning,” *Procedia Comput. Sci.*, vol. 177, pp. 618–623, 2020, doi: 10.1016/j.procs.2020.10.088.

[13] Z. Wang, Y. Zhou, T. Takagi, J. Song, Y. S. Tian, and T. Shibuya, “Genetic algorithm-based feature selection with manifold learning for cancer classification using microarray data,” *BMC Bioinformatics*, vol. 24, no. 1, pp. 1–22, 2023, doi: 10.1186/s12859-023-05267-3.

[14] H. Panwar, P. K. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh, “A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images,” *Chaos, Solitons and Fractals*, vol. 140, no. August, p. 110190, 2020, doi: 10.1016/j.chaos.2020.110190.

[15] E. Hussain, M. Hasan, M. A. Rahman, I. Lee, T. Tamanna, and M. Z. Parvez, “CoroDet: A deep learning based classification for COVID-19 detection using chest X-ray images,” *Chaos, Solitons and Fractals*, vol. 142, p. 110495, 2021, doi: 10.1016/j.chaos.2020.110495.

[16] D. Al-Karawi, S. Al-Zaidi, N. Polus, and S. Jassim, "Machine Learning Analysis of Chest CT Scan Images as a Complementary Digital Test of Coronavirus (COVID-19) Patients," *medRxiv*, no. April, 2020, doi: 10.1101/2020.04.13.20063479.

[17] S. T. Ahmed and S. M. Kadhem, "Alzheimer's disease prediction using three machine learning methods," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 27, no. 3, pp. 1689–1697, 2022, doi: 10.11591/ijeecs.v27.i3.pp1689-1697.

[18] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. Jamalipour Soufi, "Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning," *Med. Image Anal.*, vol. 65, pp. 1–9, 2020, doi: 10.1016/j.media.2020.101794.

[19] A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, "Artificial intelligence and machine learning to fight covid-19," *Physiol. Genomics*, vol. 52, no. 4, pp. 200–202, 2020, doi: 10.1152/physiolgenomics.00029.2020.

[20] D. Singh, V. Kumar, Vaishali, and M. Kaur, "Classification of COVID-19 patients from chest CT images using multi-objective differential evolution–based convolutional neural networks," *Eur. J. Clin. Microbiol. Infect. Dis.*, vol. 39, no. 7, pp. 1379–1389, 2020, doi: 10.1007/s10096-020-03901-z.

[21] E. M. Hameed, I. S. Hussein, H. G. A. Altameemi, and Q. K. Kadhim, "Liver Disease Detection and Prediction Using SVM Techniques," in *2022 3rd Information Technology To Enhance e-learning and Other Application (IT-ELA)*, Dec. 2022, pp. 61–66, doi: 10.1109/IT-ELA57378.2022.10107961.

[22] M. D. Antonio *et al.*, "Article SARS-CoV-2 susceptibility and COVID-19 disease severity are associated with genetic variants affecting gene expression in a variety of tissues ll SARS-CoV-2 susceptibility and COVID-19 disease severity are associated with genetic variants affectin," *IEEE Publ.*, 2021, doi: 10.1016/j.celrep.2021.110020.

[23] D. Li *et al.*, "COVID-19 disease and malignant cancers : The impact for the furin gene expression in susceptibility to," *Int. J. Biol. Sci.*, vol. 17, no. Cd, 2021, doi: 10.7150/ijbs.63072.

[24] N. E. M. Khalifa, M. H. N. Taha, D. Ezzat Ali, A. Slowik, and A. E. Hassanien, "Artificial intelligence technique for gene expression by tumor RNA-Seq Data: A novel optimized deep learning approach," *IEEE Access*, vol. 8, pp. 22874–22883, 2020, doi: 10.1109/ACCESS.2020.2970210.

[25] T. Burak and A. Ibrahim, "ORIGINAL RESEARCH ARTICLE A Novel Protein Mapping Method for Predicting the Protein Interactions in COVID - 19 Disease by Deep Learning," *Interdiscip. Sci. Comput. Life Sci.*, vol. 13, no. 1, pp. 44–60, 2021, doi: 10.1007/s12539-020-00405-4.

[26] M. Olaolu, A. Roseline, O. Ogundokun, S. Misra, B. Dorothy, and A. Brij, "Machine learning-based IoT system for COVID-19 epidemics," *Computing*, vol. 105, no. 4, pp. 831–847, 2023, doi: 10.1007/s00607-022-01057-6.

[27] R. Doewes, U. S. Maret, R. Nair, and T. Sharma, "Diagnosis of COVID-19 through blood sample using ensemble genetic algorithms and machine learning classi fi er," *World J. Eng.*, no. September, 2021, doi: 10.1108/WJE-03-2021-0174.

[28] A. Altan and S. Karasu, "Recognition of COVID-19 disease from X-ray images by hybrid model consisting of 2D curvelet transform, chaotic salp swarm algorithm and deep learning technique," *elsevier*, no. January, 2020.

[29] S. T. Ahmed, Q. K. Kadhim, H. S. Mahdi, and W. S. A. Almahdy, "THE USE OF SPATIAL RELATIONSHIPS AND OBJECT IDENTIFICATIO N IN IMAGE UNDERSTANDING," *Int. J. Civ. Eng. Technol.*, vol. 9, no. 5, pp. 487–496, 2018.

[30] S. T. Ahmed and S. M. Kadhem, “Early Alzheimer’s disease detection using different techniques based on microarray data: A review,” *Int. J. Online Biomed. Eng.*, vol. 18, no. 04, pp. 106–126, Mar. 2022, doi: 10.3991/ijoe.v18i04.27133.

[31] S. H. Dhahi, E. H. Dhahi, B. J. Khadhim, and S. T. Ahmed, “Using support vector machine regression to reduce cloud security risks in developing countries,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 30, no. 2, pp. 1–8, 2023, doi: 10.11591/ijeecs.v30.i2.pp1-1x.

[32] Q. K. Kadhim, “Classification of Human Skin Diseases using Data Mining,” *Int. J. Adv. Eng. Res. Sci.*, vol. 4, no. 1, 2017, doi: 10.22161/ijaers.4.1.25.

[33] Q. K. Kadhim, R. Yusof, and H. S. Mahdi, “A Review Study on Cloud Computing Issues A Review Study on Cloud Computing Issues,” *J. Phys. Conf. Ser. Pap.*

[34] X. Zeng *et al.*, “Repurpose Open Data to Discover Therapeutics for COVID-19 Using Deep Learning,” *J. Proteome Res*, 2020, doi: 10.1021/acs.jproteome.0c00316.

[35] L. Chen, Z. Mei, W. Guo, S. Ding, T. Huang, and Y. Cai, “Recognition of Immune Cell Markers of COVID-19 Severity with Machine Learning Methods,” *Hindawi Publ. Corp. BioMed*, vol. 2022, no. January 2020, 2022.

[36] J. Rokne and R. Alhajj, “A survey of machine learning-based methods for COVID-19 medical image analysis,” *Med. Biol. Eng. Comput.*, pp. 1257–1297, 2023.

[37] M. Alazab, A. Awajan, A. Mesleh, A. Abraham, V. Jatana, and S. Alhyari, “COVID-19 Prediction and Detection Using Deep Learning,” no. June, 2020.

[38] I. D. Apostolopoulos and T. A. Mpesiana, “Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks,” *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 635–640, 2020, doi: 10.1007/s13246-020-00865-4.

[39] S.-W. Hao-Yuan Cheng, “Contact Tracing Assessment of COVID-19 Transmission Dynamics in Taiwan and Risk at Different Exposure Periods Before and After Symptom Onset,” vol. 180, no. 9, pp. 1156–1163, 2020, doi: 10.1001/jamainternmed.2020.2020.

[40] A. Bansal, R. P. Padappayil, C. Garg, A. Singal, M. Gupta, and A. Klein, “Utility of Artificial Intelligence Amidst the COVID 19 Pandemic :,” *J. Med. Syst.*, 2020.

[41] R. G. Babukarthik, V. A. K. Adiga, and G. Sambasivam, “Prediction of COVID-19 Using Genetic Deep Learning Convolutional Neural Network (GDCNN),” *IEEE Access*, vol. 8, pp. 177647–177666, 2020, doi: 10.1109/ACCESS.2020.3025164.

[42] T. Therapy, “Drug repurposing for COVID-19 using machine learning and mechanistic models of signal transduction circuits related to SARS-CoV-2 infection,” *Signal Transduct. Target. Ther.*, no. November, pp. 19–21, 2020, doi: 10.1038/s41392-020-00417-y.

[43] B. K. Patterson *et al.*, “Immune-Based Prediction of COVID-19 Severity and Chronicity Decoded Using Machine Learning,” *Front. Pharmacol.*, vol. 12, no. June, pp. 1–13, 2021, doi: 10.3389/fimmu.2021.700782.

[44] M. Kavitha, T. Jayasankar, P. M. Venkatesh, G. Mani, C. Bharatiraja, and B. Twala, “COVID-19 Disease Diagnosis using Smart Deep Learning Techniques,” *J. ofApplied Sci. Eng.*, vol. 24, no. 3, pp. 271–277, 2020.

[45] M. Hasan, S. B. Murtaz, M. U. Islam, J. Sadeq, and J. U. Id, “Robust and efficient COVID-19 detection techniques : A machine learning approach,” *PLOSE ONE*, vol. 2, pp. 1–21, 2022, doi: 10.1371/journal.pone.0274538.

[46] A. Alimadadi, S. Aryal, I. Manandhar, X. P. B. Munroe, and B. Joe, “Artificial intelligence and machine learning to fight COVID-19,” *Physiol Genomics 52*, pp. 200–202, 2023, doi: 10.1152/physiolgenomics.00029.2020.