World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:15, No:6, 2021

Comparative Analysis of Machine Learning Tools: A Review

S. Sarumathi, M. Vaishnavi, S. Geetha, P. Ranjetha

***Abstract*—**Machine learning is a new and exciting area of artificial intelligence nowadays. Machine learning is the most valuable, time, supervised, and cost-effective approach. It is not a narrow learning approach; it also includes a wide range of methods and techniques that can be applied to a wide range of complex real-world problems and time domains. Biological image classification, adaptive testing, computer vision, natural language processing, object detection, cancer detection, face recognition, handwriting recognition, speech recognition, and many other applications of machine learning are widely used in research, industry, and government. Every day, more data are generated, and conventional machine learning techniques are becoming obsolete as users move to distributed and real-time operations. By providing fundamental knowledge of machine learning tools and research opportunities in the field, the aim of this article is to serve as both a comprehensive overview and a guide. A diverse set of machine learning resources is demonstrated and contrasted with the key features in this survey.

***Keywords*—**Artificial intelligence, machine learning, deep learning, machine learning algorithms, machine learning tools.

I. Introduction

IN many research fields such as deep learning, pattern matching, information retrieval, medical fields, image processing, spatial data retrieval, industry, and education, the domain of machine learning and extracting useful, novel, and true patterns exploded in recent years [1], [2]. Machine learning's purpose is to allow a system to interpret information from the past or present and use that information to make predictions or decisions about known and unknown future class labels. Learning refers to the process of extracting useful patterns from unstructured data. The three types of algorithms used in machine learning are supervised learning, unsupervised learning, and reinforcement learning. Patterns are inferred from the labelled input file in supervised learning, which aids in the prediction of outcomes from unexpected results. Fig. 1 depicts the supervised machine learning workflow. Building the model, evaluating and tuning the model, and then deploying the model prediction are all part of the supervised machine learning workflow. Classification and regression problems are common categories for supervised learning. Rectilinear regression, logistic regression, help vector machine, multi-class classification, decision tree, Bayesian logic, and other algorithms are also included in machine learning. Unsupervised learning is a machine learning approach that involves inferring data from unlabeled input file patterns. Clustering and association issues are common classifications for unsupervised learning. Clustering, k-nearest neighbor, and the apriori algorithm are among the algorithms used. Different software and machines use Reinforcement Learning (RL) to find the simplest possible action or direction depending on the current situation. RL solves a specific type of problem where the decision-making is linear and the aim is long-term, such as game-playing and robotics. To extract accurate patterns and trends from the data, this machine learning model relies on complex algorithms and mathematical analysis [3].

The main goal of machine learning is to construct a useful predictive and descriptive model out of a large amount of data that is collected in real time. Several real-world machine learning problems contain many inconsistencies in output or target, all of which must be optimized at the same time. One of the most distinguishing characteristics of machine learning is that it deals with large and complex datasets ranging in size from gigabytes to terabytes. This necessitates stable, scalable machine learning tools and algorithms, as well as the ability to collaborate across various research domains. As a result, various machine learning tools play an important role in each and every aspect of data extraction, resulting in the advent of many deep learning tools. From a functional standpoint, the graphical interfaces used in the tools are more powerful, user-friendly, and simpler to work with, which is why researchers prefer them. Fig. 2 depicts the machine learning's traditional basic structure.

The following is how the rest of this paper will be organized: Section II provides an overview that can be applied to real-world problems. Section III gives a brief overview of the tools and their key features. Section IV gives an overview of the tools that can be used to solve real-world problems. Section V examines different parameters, highlights features, and provides a brief overview of specific frameworks that can be used with the processing platforms. In the final section, the conclusions of the survey are discussed.

II. Popular Machine Learning Algorithms

Some popular and most commonly used machine learning algorithms along with their use cases and applications are as follows [4]-[8]:

- Linear Regression algorithm
- Logistic Regression algorithm
- Decision Tree algorithm
- Support Vector Machine algorithm
- Naïve Bayes algorithm

Dr. S. Sarumathi, Professor, M.Vaishnavi, Assistant Professor, Ms. S. Geetha, Assistant Professor, and Ms. P. Ranjetha, Assistant Professor, are with the Department of Information Technology, K. S. Rangasamy College of Technology, Tamil Nadu, India (e-mail: rishi_saru20@rediffmail.com, vaishnavi.munusamy@gmail.com,geethas@ksrct.ac.in, ranjetha@ksrct.ac.in).

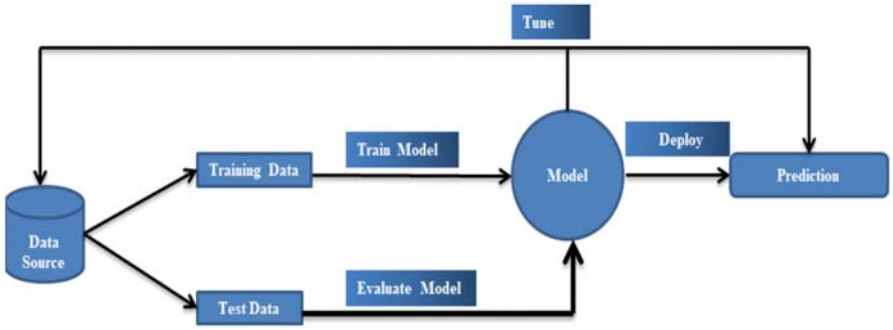

Fig. 1 Supervised machine learning workflow

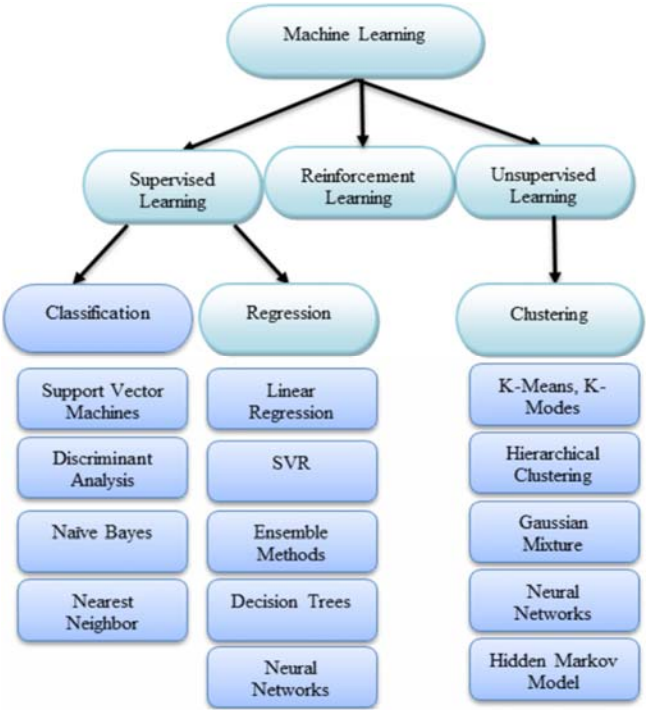

Fig. 2 Basic Process in Machine Learning

- k-Nearest Neighbor algorithm
- K-Means Clustering algorithm
- Random Forest algorithm
- Apriori algorithm
- Principal Component Analysis algorithm

A. Linear Regression Algorithm

Linear regression is a basic machine learning technique that is primarily used for predictive analysis. Linear regression is used to forecast continuous numbers such as age, income, and grades, among other things. There are two different forms of linear regression:

- *Simple Linear Regression:* This is a technique for predicting the value of a dependent variable using only one independent variable.
- *Multiple Linear Regression:* This is a method where more than one independent variable are used to predict the value of the dependent variable.

Risk assessment in the financial services or insurance domain, econometrics, epidemiology, weather data analysis, predictive analytics, and customer survey results analysis are some examples of real-world applications of linear regression.

B. Logistic Regression Algorithm

Another supervised learning algorithm is logistic regression, which is used to predict categorical variables or discrete values. It is primarily used in machine learning for classification problems, and the performance of the logistic regression algorithm can be yes or no, 0 or 1, red or blue, and so on.

Logistic regression is similar to linear regression, except that logistic regression is used to solve the classification problem and predict discrete values, whereas linear regression is used to solve the regression problem and predict continuous values.

Trauma and injury severity scores, image segmentation and categorization, cancer identification, geographic image analysis, handwriting recognition, and prediction are some examples of real-world applications of logistic regression.

C. Decision Tree Algorithm

A decision tree is a supervised learning algorithm that is widely used to solve the classification problem. It is used to solve regression problems and can handle both categorical and continuous variables. It depicts a tree-like structure with nodes and branches, beginning with the root node and progressing through the branches to the leaf node. The internal node represents the dataset's characteristics, while branches represent decision rules and leaf nodes represent the problem's outcome.

Decision tree algorithms are used to classify cancerous and non-cancerous cells, as well as make recommendations to customers looking to purchase a vehicle.

D. Support Vector Machine Algorithm

Another supervised learning algorithm for classification and regression problems is the Support Vector Machine (SVM). It is, however, widely used to solve classification problems. SVM's goal is to create a hyperplane or decision boundary that can divide datasets into distinct groups.

Support vectors are the data structures that help describe the hyperplane, so the algorithm is called a SVM algorithm. SVM has real-world applications in face recognition, image classification, and drug discovery.

E. Naïve Bayes Algorithm

The Naive Bayes classifier is a supervised learning algorithm that makes predictions based on the object's probability. The Bayes algorithm is used in the Naive Bayes algorithm, which is based on the assumption that variables are unrelated.

One of the most powerful classifiers for solving a known problem is the Naive Bayes classifier. A Naive Bayesian model is simple to construct and is well suited to large data sets. It is often used in text classification.

F. k-Nearest Neighbor (k-NN) Algorithm

The k-Nearest Neighbor algorithm is classified as a supervised learning algorithm that can be used to solve classification and regression problems. The k-NN algorithm works by assuming that new data objects and existing data objects are identical. Based on these similarities, the new data objects are placed in the most similar groups. It is also known as the lazy learner algorithm because it stores all of the available data sets and uses k-neighbors to classify each new case. Based on the distance between the data objects, the new data are assigned to the closest class with the most similarities. Based on the criterion, Euclidean, Minkowski, Manhattan, or Hamming distance are widely used distance functions.

The k-NN is most commonly used in accounting, credit scoring, speech recognition, handwriting recognition, image recognition, and video recognition.

G. K-Means Clustering Algorithm

In partitioning clustering, K-means clustering is one of the simplest unsupervised learning algorithms for resolving clustering problems. Based on similarities and dissimilarities, the data objects are grouped into k distinct clusters that are mutually disjoint. The number of clusters in K-means clustering is k, and the means denotes the averaging of data points to find the centroid.

The k-means clustering algorithm primarily performs two functions. Iteratively decides the best value for K center points or centroids in the first step. It then assigns each data point to the k-center that is nearest to it. It is mainly used for detecting and filtering spam, as well as identifying fake news.

H. Random Forest Algorithm

Random forest is a supervised learning algorithm used in machine learning. This algorithm is used in machine learning for both classification and regression problems. It makes predictions by combining several classifiers and improving the model's accuracy using an ensemble learning technique. It is made up of multiple decision trees for different subsets of a dataset, with an average estimated to improve the model's predictive accuracy. The number of trees in a random forest must be between 64 and 128. As the number of trees grows, the algorithm's accuracy improves automatically.

Random forest is a fast algorithm that can deal effectively with missing and incorrect data. Random forest algorithm has applications in the banking industry, credit card fraud detection, and customer segmentation.

I. Apriori Algorithm

The apriori algorithm is a type of unsupervised learning algorithm that is used to find frequent itemsets in transactional datasets. It is mostly used to solve problems involving associations. Association rules are created from frequent itemsets, which decide how strongly or weakly two objects are related to each other. A breadth-first search and FP growth are used in this algorithm.

The algorithm iterates through the huge transactional dataset to find the frequent itemsets. It is most commonly used in market basket research to help classify items that can be purchased together. It is also used in medicine to detect drug reactions in patients.

J. Principal Component Analysis Algorithm

Principal Component Analysis (PCA) is an unsupervised learning technique that is primarily used to reduce dimensionality. It decreases the dimensionality of a dataset with several attributes that are similar to one another. It uses orthogonal transformation to convert observations of correlated features into a collection of linear uncorrelated features. This program is used for data exploration and predictive modelling. The variance of each attribute is checked in PCA since a high variance indicates a good split between the groups and thus reduces dimensionality.

In domains like facial recognition, computer vision, and image compression, PCA is primarily used as a dimensionality reduction technique.

III. Open Source Machine Learning Tools

The following are the five categories of common open source machine learning tools [9]:

A. Open Source Machine Learning Tools for Non-Programmers

- *Uber Ludwig:* This is a toolbox that allows you to coach and evaluate deep learning models without writing code by providing a CSV file with a list of columns to use as inputs and a set of columns to use as outputs. Ludwig takes care of the rest of the operation, which includes training and testing models.
- *KNIME:* This programme is primarily used to design and incorporate data science algorithms in a visual workflow model. It has a drag-and-drop gui that makes complicated problem statements simple.
- *Orange:* This is a machine learning, data visualization, and data mining toolkit that is open-source. It comes with a visual programming front-end for exploratory data analysis and immersive data visualization, as well as the ability to be used as a Python library. It has bioinformatics and text mining add-ons.

B. Machine Learning Model Deployment

- *MLFlow:* This is a machine learning development framework that provides a collection of lightweight APIs that can be integrated into any existing machine learning application or library. It is in charge of keeping track of

tests, packaging machine learning code, and handling and deploying models.

- *Apple's CoreML:* CoreML has unified representation for all the models. It supports analyzing images, text processing and sound analysis. CoreML is an apple framework to integrate machine learning models into an app.
- *TensorFlow Lite:* This is a tool that allows developers to train and infer deep neural network models that run on mobile devices such as Android and iOS, as well as IoT devices and embedded systems. The binary size and output of both the TenserFlow Lite interpreter and TenserFlow Lite converter have been designed for many hardware models.
- *TensorFlow.js:* This is an open source library that allows users to deploy machine learning models on the web. It aids in the creation of machine learning models and even runs alongside existing models. It is also used to retrain existing machine learning models with our own results.

C. Big Data Open Source Tools

- *Hadoop:* This is capable of storing and processing various types of massive data in a distributed environment. It is an open source platform that allows you to scale from a single server to a cluster of machines. It includes computation and storage on the local computer as well.
- *Spark:* Apache Spark is an open source data processing engine that makes analytics for batch and real-time data simpler and easier to use.
- *Neo4j:* This is a graph database that deals with real-world data and the relationships between it. According to the query, data modelling focuses and retrieves information in an effective manner.

C. Open Source Machine Learning Tools for Computer Vision, NLP, and Audio

- *SimpleCV:* This is a computer vision programme that is free and open source. It is beneficial for beginners to learn how to perform machine vision tests quickly. In simpleCV, manipulations happen very quickly.
- *Tesseract OCR:* This is primarily concerned with line recognition. It comes preloaded with over 100 languages and can be used to train other languages as well. It is primarily used in text detection, video detection, and Gmail picture spam detection on mobile devices.
- *Detectron:* This is a Facebook AI analysis software framework. It is derived from Mask R-CNN and is primarily used to implement object detection algorithms. The PyTorch deep learning system trains the data set much faster.
- *StanfordNLP:* This is a powerful deep linguistic modelling and data analysis software kit. It now supports over 70 human languages as a result of effective training and evaluation. Languages are transformed into strings that contain text. A syntactic structure dependency parse is obtained by converting the text into a list of sentences and phrases.
- *BERT as a Service:* This is a sentence encoding service that is highly scalable. It converts a variable-length sentence into a vector of fixed length (i.e., only two lines of code). It is one of the pre-train models, and it was created by Google to train data found on the internet.
  Google Magenta: The Google Brain team created Google Magenta, which uses deep learning and RL algorithms to train music and photos into new content.
- *Librosa:* For analyzing video and music files, the Librosa Python package is used. It has structure building blocks for retrieving details. It's a submodule-based framework. It includes a massive amount of audio signal preprocessing in order to perform deep learning-based audio to text conversion.

D. Open Source Tools for RL

- *Google Research Football:* This is an RL-inspired 3D football game. A well-optimized game engine teaches us how to pass the ball and play defense, as well as how to control one or more players on the squad.
- *OpenAI Gym:* This is a modular collection of environments for designing and comparing RL algorithms, which aids us in learning new hardcoded game solvers and deep learning approaches. In the robot simulation, the designs are extremely precise and fast.
- *Unity ML Agents:* The Unity Machine Learning Agent Toolkit (ML-Agents) trains intelligent agents for 2D, 3D, and VR/AR games to allow games and simulations. For both game development and AI analysis techniques, the Python API is used to train the agents.
- *Project Malmo:* It is a sophisticated AI experimental platform designed on top of Minecraft to support AI fundamental research. It allows you to look at and edit the Minecraft code.

Fig. 3 depicts a general classification of machine learning tools.

IV. Different Machine Learning Tools

A. Tensorflow

Tensorflow provides a JavaScript library for deep learning and machine learning applications, as well as APIs to help with data acquisition, model testing, serving predictions, and optimizing potential results [10]. TensorFlow is an open source software library or platform created by the Google team to simplify the execution of machine learning and deep learning concepts. It includes a range of machine learning and deep learning algorithms, as well as optimization techniques [11]-[14]. It also includes a variety of machine learning and deep learning algorithms for simple computation of several mathematical expressions. Tensorflow is used to train and run deep neural networks for image recognition, handwritten digit classification, word embedding, and the construction of various sequence models.

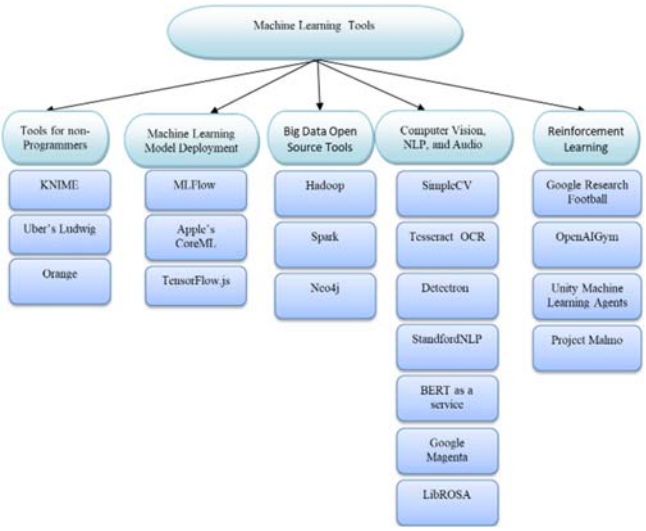

Fig. 3 Classification of Machine Learning Tools

Pros:
- A user can run existing models using TensorFlow.js, a model converter.
- The neural network can be used with Tensorflow.
- Python tools allow for faster debugging.
- Python controls flow for dynamic models
- Custom and higher-order gradients are supported.
- TensorFlow allows to create and train models at multiple levels of abstraction.
- TensorFlow allows users to easily train and deploy models, regardless of language or platform.
- TensorFlow features like the Keras Functional API and Model provide versatility and power.
- It is well-documented and simple to comprehend
- It is perhaps the most widely used Python package.

Cons:
- Learning is difficult.
- Other than Nvidia, there is no GPU support.
- Computation speed is fast

B. Keras.io

Keras is a human-oriented API, not a machine-oriented API. Keras offers reliable and straightforward APIs, as well as transparent and actionable error messages and detailed documentation and developer guides. Keras was used by CERN, NASA, NIH, and several other research organizations around the world [15], [16]. Keras offers high-level convenience features that are versatile enough to execute arbitrary research ideas to speed up experimental cycles.

Pros:
- The Keras functional API can be used to create multi-input/multi-output models, directed acyclic graphs (DAGs), and models with shared layers.
- It helps users to execute their activities with the least amount of effort possible.
- The use of convolutional networks is encouraged.
- The use of recurrent networks is aided.
- It operates for a variety of networks.
- It uses a single GPU to train a model or multiple GPUs to train a model.
- It is only accessible in Python and does not support any other languages.
- It is designed with the goal of allowing for fast experimentation.

Cons:
- TensorFlow, Theano, or CNTK are needed to use Keras.

C. PyTORCH

PyTorch is a machine learning library focused on Torch for deep learning on GPUs and CPUs [17]. The torch is a Lua-based computing platform, scripting language, and machine learning library that replaces NumPy to take advantage of GPU computing resources. PyTorch is the system of choice for a large number of researchers, engineers, and developers since it is mainly used to train deep learning models quickly and efficiently.

Pros:
- The Autograd Module is used to build neural networks.
- It has a number of optimization algorithms for building
- neural networks.
- It can be found on cloud services.
- It provides online instruction, as well as different resources and libraries.
- It assists with the development of computational graphs.
- User-friendliness refers to the hybrid front-end.

Cons:
- It is recent and not well-known.
- It lacks model serving in production.
- Tensorboard, for example, lacks control and visualization interfaces.

D. Shogun

Shogun offers a wide range of machine learning algorithms and data structures. These libraries for machine learning are used in education and science [18]-[20].

Pros:
- Vector machines may be used for regression and classification.
- It assists in the implementation of Hidden Markov models.
- It supports a variety of languages, including Python, Octave, R, Ruby, Java, Scala, and Lua.
- It works with massive datasets.
- It is easy to use.
- It provides excellent customer service.
- It provides useful features and functions.

E. Apache Mahout

This is a machine learning framework that allows developers to build scalable machine learning algorithms. Mathematicians, statisticians, and data scientists may use Apache Mahout to run their algorithms [21], [22].

Pros:
- Pre-processing, regression, clustering, recommenders, and distributed linear algebra algorithms are all included.
- The Java libraries contain popular math operations.

- It uses a structure for distributed linear algebra.
- It is suitable for massive data sets.
- It is easier and faster to build intelligent applications.
- It makes use of the Apache Hadoop library to help Mahout scale more quickly in the cloud.

Cons:

- More helpful documentation is needed.
- There are several algorithms that are incomplete.

F. Accord.Net

Accord.NET is a.NET framework that supports scientific computing. It is made up of several libraries that cover a broad variety of scientific computing applications, including but not limited to machine learning, pattern recognition, statistical data analysis, computer vision, and computer audition [23]. This framework also includes a large number of probability distributions, kernel functions, hypothesis tests, and support for the majority of commonly used performance assessment techniques.

Pros:

- Algorithms for numerical linear algebra, numerical optimization, statistics, and artificial neural networks are all available.
- It provides image, audio, and signal processing machine
- learning libraries.
- Chart plotting and visualisation libraries are fully supported.
- Libraries are included in the source code, as well as an executable installer and a NuGet package manager.

Cons:

- It only works for .Net Languages that are supported on the internet.

G.Rapidminer

This is a data mining software application that is both open source and commercial [24]. It has been implemented in a framework that includes data mining, text mining, machine learning, deep learning, business analytics, and predictive analytics. The rapidminer platform is a stand-alone framework for combining data analysis and data mining engine products. As used in more than 40 countries, it gives users competitive advantages in their applications [25].

Pros:

- Analytical methods are designed and executed using graphical user interfaces.
- It can be used to prepare data.
- Data finding and visualization tools are simple to use^.
- Model validation and optimization are quick and easy.
- Extensions may be used to extend the functionality.
- It is easy to use.
- No programming skills are required.

Cons:

- The tool is expensive.

H. Weka

Weka is an open source tool with a graphical user interface, regular terminal programs, and a Java API [26]-[28]. It is widely used in science, industry, and education, and it comes with a number of built-in tools for popular machine learning tasks, as well as easy access to well-known toolboxes like scikit-learn, R, and Deeplearning4j. The Deeplearning4j package was created to incorporate current deep learning techniques into Weka [29]. Without requiring the use of code, it assists in the development of machine learning pipelines, the training of classifiers, and the execution of evaluations.

Pros:

- Algorithms are simple to comprehend.
- It is also beneficial to students.
- It offers instruction through online courses.

Cons:

- There is no documentation or online assistance available.

I. Google Colab

Google Colab is a cloud service that supports Python and can be used to create machine learning applications using PyTorch, Keras, TensorFlow, and OpenCV libraries. Google Colab is a fully cloud-based Jupyter notebook environment [30].

Pros:

- It aids machine learning education.
- It aids machine learning research.
- Colab can be used from Google Drive.

Cons:

- All specific libraries must be installed.

J. Scikit-Learn

Scikit-learn is primarily used in machine learning and python development. It is a Python library. It also provides a clear interface in Python for a variety of supervised and unsupervised learning algorithms [31].

Pros:

- It aids in data mining and analysis.
- It provides models and algorithms for classification, regression, clustering, dimensional reduction, model selection, and pre-processing.
- Parameters for any algorithm can be modified when calling artefacts.

Cons:

- In modelling, the function is biased.

K. R Studio

R is extremely extensible and offers a broad range of statistical techniques, including linear and nonlinear modelling, classical statistical studies, time-series analysis, classification, clustering, and graphical techniques. R programming applications range from hypothetical to computational statistics and the hard sciences, such as physics, chemistry, and genomics, to practical applications in industry, drug development, finance, health care, marketing, and medicine, among many others. Nearly 5,000 packages are available in R. (libraries of functions). Many quantitative analysts in finance use R as their primary programming method [32].

Pros:

- R is an open-source language without any need for a license

- Its code can run on all operating systems.
- It provides various packages and features for developing the artificial neural network
- Data wrangling feature in R transforms messy data into a structured form.
- R simplifies quality plotting and graphing
- R has over 10000 packages for data science and machine learning operations
- R is predominant than other programming languages for the development of statistical tools.
- R is a constantly evolving programming language

Cons:

- It does not allow dynamic or three-dimensional graphics.
- As compared to Python, R uses more memory.
- R is not suitable for use in a web application.
- Learning R without prior experience is difficult.
- R packages are significantly slower.

L. Oryx 2

For real-time and large-scale machine learning analysis, Oryx 2 employs Lambda Architecture. This model was created using the Apache Spark architecture, which consists of bundled functions for rapid prototyping and application development. It simplifies the development of end-to-end models for collaborative filtering, classification, regression, and clustering operations [33].

The three levels of Oryx 2 are as follows.

- The lambda tier, which offers speed and serving layers that are not unique to Machine Learning procedures, is the first tier.
- The second-tier specialization provides ML abstractions for hyperparameter selection.
- The same standard ML algorithms (ALS, random decision forests, k-means) are utilized as an application in the top tier.

Pros:

- Oryx 2 is a three-tiered machine learning system that focuses on real-time, large-scale machine learning.

Cons:

- It doesn't have a large number of algorithms to choose from.

M. H20.ai

The H20 deep learning framework includes a multi-layer artificial neural network that is scalable. It is a distributed in-memory machine learning framework with linear scalability that is entirely open source. It supports a variety of statistical and machine learning algorithms, such as gradient boosted machines, generalized linear models, deep learning, and others. The components and parameters of this ANN can be changed depending on the data given. Convolutional and Recurrent Neural Networks are also supported [34].

Pros:

- Leading Algorithms
- Access from R and Python, etc.,
- AutoML
- Distributed, In-Memory Processing Simple Deployment

N. Orange

Orange is a C++ core object and routines library that supports a wide range of machine learning and data mining algorithms [30]. Orange is a GPL-licensed open-source software package that provides a rich collection of mining and machine learning algorithms for data pre-processing, classification, modelling, regression, clustering, and other functions in Python scripts.

Orange's goal is to serve as a forum for experiment-based selection, predictive modelling, and recommendation systems. It is mostly used in bioinformatics, genomics, biomedicine, and education. It is used in education to provide improved data mining and machine learning teaching methods for students of genetics, biomedicine, and informatics [35]. Orange also includes a visual programming environment, with a workbench that includes tools for importing data, dragging and falling widgets, and linking various widgets together to complete the workflow. Orange widgets include a graphical user interface for data mining and machine learning techniques used by Orange. They include widgets for data entry and preprocessing, classification, regression, association rules, and clustering, as well as a collection of widgets for model evaluation and visualization of assessment outcomes, as well as widgets for exporting models to PMML, as shown in Fig. 4.

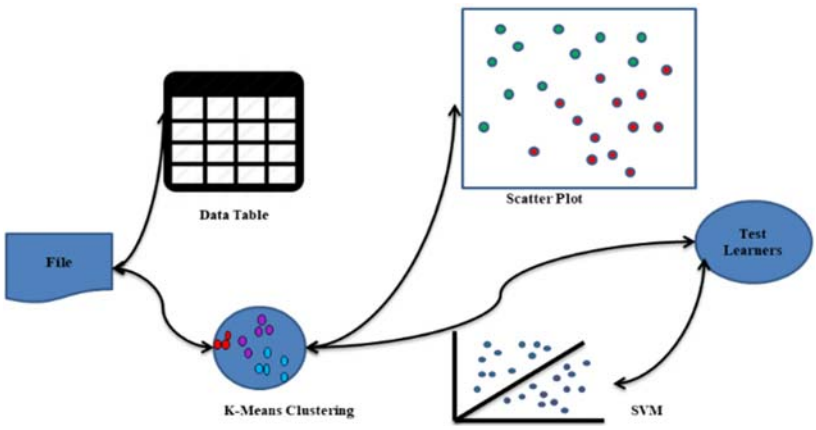

Fig. 4 Orange Tool Widget Flow

TABLE I
COMPARISON OF MACHINE LEARNING TOOLS

Name of the Tool	Platform Supported	Mode of Software	Written Language	Features	Applications
Tensorflow	Linux, MacOS, Windows	Free	Python, C++, CUDA	-Provides a Library for Dataflow Programming.	Speech Recognition Models, Predictive Analytics, Video Detection, Text-Baaed Applications
Keras.io	Cross-platform	Free	Python	-API for Neural Networks	Deep Learning, Prediction
Shogun	Windows Linux UNIX Mac OS	Free	C++	-Regression -Classification -Clustering - SVMs -Dimensionality Reduction -Online Learning	Machine Learning, Bioinformatics
Apache Mahout	Cross-platform	Free	Java Scala	-Preprocessing -Regression -Clustering -Recommenders -Distributed Linear Algebra	Pattern Mining, Recommender Systems, Modeling and Prediction
Accord.Net	Cross-platform	Free	C#	-Classification -Regression -Distribution -Clustering -Hypothesis Tests -Kernel Methods -Image Vision -Audio Vision	Statistical Data Processing, Machine Learning, Pattern Recognition, Computer Vision, Computer Audition.
PyTorch	Linux, Mac OS, Windows	Free	Python, C++, CUDA	-Autograd Module -Optim Module -nn Module	Image Classification, Handwriting Recognition, Forecast Time Sequences, Text Generation, Style Transfer,
Rapid Miner	Cross-platform	Free and Commercial	Java	-Data Loading -Data Transformation -Data Preprocessing -Data Visualization	Prediction, Sentiment Analysis, Fraud Detection, Healthcare
Weka	Linux, Mac OS, Windows	Free	Java	-Data Preparation -Classification -Regression -Clustering -Visualization -Association Rule Mining	Predictive and Descriptive Modeling, Data Visualization, Deep Learning
Colab	Cloud Service	Free	-Python	-Support Libraries of PyTorch, Keras, TensorFlow, and OpenCV -Free cloud service -Support mathematical equations	Deep Learning, GPU Centric Applications
Scikit Learn	Linux, Mac OS, Windows	Free	Python, Cython, C, C++	-Classification -Regression -Clustering -Preprocessing -Model Selection -Dimensionality Reduction	Financial Cyber Security Analytics, Product Development, Neuroimaging, Barcode Scanner Development, Medical Modeling
R	Linux, MacOS, and Windows	Free	C	-Classification -Regression -Clustering -Preprocessing -Model Selection -Dimensionality Reduction	Statistics, data analysis, and machine learning
Oryx 2	Cross-platform	Free	Java	-Classification -Regression -Clustering -Filtering	Collaborative filtering, classification, regression and clustering.
H2O.ai	Open Source platform	Free	Java	-Classification -Prediction	Artificial Intelligence, Machine Learning
Orange	Cross Platform	Free	Python, Cython, C++, C	Classification Data Visualization	visual programming front-end for explorative rapid qualitative, data analysis and interactive data visualization

Tokens are transferred from the sender widget to the receiver widget to convey data. For scientific computing, Orange relies on open-source Python libraries such as numpy, scipy, and scikit-learn, while its graphical user interface is built on the cross-platform Qt framework [36].

Pros:

- The Orange tool works with the NumPy and SciPy libraries, as well as reading online info.
- Reading online data, working through SQL queries, and pre-processing are some of Orange's other enhancements.
- It assists you in predicting market insights through machine learning.

Cons:

- Graphics in their raw form cannot be copied or pasted into any presentation.
- In green, live data analysis and prediction are not possible.
- Errors are not classified or checked by numbers.

V. Summarization of Machine Learning Tools

Table I shows a comparison of different machine learning tools based on their compatibility characteristics and application domains. The main goal of this comparison is not to determine which machine learning method is the best, but to demonstrate the usage model and understanding of tools in a variety of fields.

VI. Conclusion

In this paper, a number of machine learning tools were discussed, as well as how they were applied to different tasks. Each subtask of machine learning appears to be a critical reinforcement process for effective knowledge extraction. This criterion paves the way for the development of several other machine learning resources. These tools have a comprehensive technological paradigm, an excellent graphical interface, and built-in multipart algorithms that make them extremely useful for managing large amounts of data more precisely and legibly. Thus, the primary goal of this survey is to provide readers with more information about machine learning tools and their applications in various industries, which will be very useful to them, as well as to address the needs of machine learning researchers to develop more advanced tools in the future.

References

[1] Muhammad Imran Razzak, Saeeda Naz and Ahmad Zaib, 'Deep Learning for Medical Image Processing: Overview, Challenges and Future', Deep Learning for Medical Imaging, pp. 323– 350, 2017.
[2] P V Rajaraman, M Prakash, 'Deepreply - An Automatic Email Reply System with Unsupervised Cloze Translation And Deep Learning', ICTACT Journal on Soft Computing, pp. 2090 – 2095, 2020.
[3] Michael J. Bianco, Peter Gerstoft, James Traer, EmmaOzanich, Marie A. Roch, Sharon Gannot, Charles-AlbanDeledalle, 'Machine learning in acoustics: Theory and applications', The Journal of the Acoustical Society of America, 2019.
[4] https://www.javatpoint.com/machine-learning-algorithms
[5] https://www.dataquest.io/blog/top-10-machine-learning-algorithms-for-beginners/
[6] https://towardsdatascience.com/top-10-algorithms-for-machine-learning-beginners-149374935f3c
[7] Mr. Chintu Kumar, "Machine Learning Concept, Algorithms and Applications: A Survey", International Journal of Advance Research and Innovative Ideas in Education, pp.901-907,2020.
[8] Sunpreet Kaur, Sonika Jindal, "A Survey on Machine Learning Algorithms", International Journal of Innovative Research in Advanced Engineering, pp.6-14,2016
[9] https://www.analyticsvidhya.com/blog/2019/07/21-open-source-machine - learning - tools/
[10] Nguyen, G., Dlugolinsky, S., Bobák, M. et al., 'Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey'. Artif Intell Rev , pp.77–124,2019.
[11] TensorFlow, https://www.tensorflow.org/. ,2018.
[12] TF, https://www.tensorflow.org/community/roadmap,2018.
[13] TensorFlowLite,https://www.tensorflow.org/mobile/, 2018.
[14] https://tensorlayer.readthedocs.io/en/latest/ 2018.
[15] https://www.infoworld.com/article/3336192/what-is-keras-the-deep-neural-network-api-explained.html
[16] https://www.datacamp.com/community/tutorials/deep-learning-python
[17] O. Obulesu, M. Mahendra and M. ThrilokReddy, 'Machine Learning Techniques and Tools: A Survey', International Conference on Inventive Research in Computing Applications (ICIRCA),pp. 605-611, 2018.
[18] S Sonnenburg, 'The SHOGUN Machine Learning Toolbox', Journal of Machine Learning Research,pp. 1799-1802, 2010.
[19] Benjamin Hillmann, Gabriel A Al-Ghalith, Robin R Shields-Cutler, Qiyun Zhu, Rob Knight, Dan Knights, 'SHOGUN: a modular, accurate and scalable framework for microbiome quantification', Bioinformatics, pp.4088-4090,2020.
[20] S. Sarumathi, N. Shanthi, 'Comprehensive Analysis of Data Mining Tools' International Journal of Computer and Information Engineering, pp. 837-847,2015
[21] https://www.opensourceforu.com/2017/11/implementing-scalable-high-performance-machine-learning-algorithms-using-apache-mahout/
[22] https://www.infoq.com/news/2009/04/mahout/
[23] https://www.h3abionet.org/images/Technical_guides/MachineLearning_Tools_Handbook_ML_project.pdf
[24] S. Sarumathi, N. Shanthi, S. Vidhya, M. Sharmila,' A Review: Comparative Study of Diverse Collection of Data Mining Tools', International Journal of Computer and Information Engineering, pp. 1028-1033, 2014.
[25] Rapid Miner (Online). Available at: http://www.rapidi.com/downloads/tutorial/rapidminer-4.6-tutorial.pdf.
[26] Kulwinder Kaur, Shivani Dhiman, 'Review of Data Mining with Weka Tool', International Journal of Computer Sciences and Engineering, pp.41-44, 2016.
[27] Weka (Online). Available at: http://www.gtbit.org/downloads/dwdmsem6/dwdmsem6lman.pdf
[28] Weka (Online). Available at: http://www.cs.ccsu.edu/nmarkov/weka.tutorial.pdf
[29] https://www.cs.waikato.ac.nz/ml/weka/
[30] T. Carneiro, R. V. Medeiros Da NóBrega, T. Nepomuceno, G. Bian, V. H. C. De Albuquerque and P. P. R. Filho, 'Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications', pp. 61677-61685, 2018.
[31] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort. Et. Al., 'Scikit-Learn: Machine Learning in Python', Journal of Machine Learning Research',2012.
[32] https://data-flair.training/blogs/pros-and-cons-of-r-programming-language/
[33] Sara Landset, Taghi M. Khoshgoftaar, Aaron N. Richter, and Tawfiq Hasanin, 'A survey of open source tools for machine learning with big data in the Hadoop ecosystem',Journal of Big Data,2015.
[34] https://www.h2o.ai/products/h2o-automl/
[35] https://www.javatpoint.com/orange-data-mining
[36] https://www.zeolearn.com/magazine/building-machine-learning-model-is-fun-using-orange

Sarumathi Sengottaian received B.E degree in Electronics and Communication Engineering from Madras University, Madras, Tamil Nadu India in 1994 and the M.E degree in Computer Science and Engineering from K.S. Rangasamy College of Technology, Namakkal, Tamil Nadu, India in 2007. She has completed Ph.D degree in Anna University, Chennai in Data Mining. She has a teaching experience of about 22.10 years. At present she is working as Professor in Information Technology department at K.S. Rangasamy College of technology. She has published 19 papers in the reputed International Journals and 2 papers in the reputed National journals. And also she has presented papers in five International conferences and four national conferences. She has received many cash awards for producing cent percent results in university examination. She is a life member of ISTE.

Vaishnavi Munusamy completed B.E degree in Computer Science and Engineering from Anna University, Thiruchirapalli in the year of 2011 then she got M.Tech degree in Database Systems from SRM University, Chennai in the year of 2013. She has a teaching experience around 5.6 years. Currently, she is working as an Assistant Professor in Department of Information Technology at K.S. Rangasamy College of technology. She has presented 2 papers in International Conference. She has received many cash awards for producing cent percent results in university examination.

S. Geetha holds a B.Tech degree in Information Technology from K.S. Rangasamy College of Technology, affiliated to Anna, University, Chennai, Tamil Nadu, India in 2014 and M.Tech degree in Information Technology from K.S. Rangasamy College of Technology, affiliated to Anna University, Chennai, Tamil Nadu, India in 2016. Now she is working as Assistant Professor in Information Technology department at K.S. Rangasamy College of Technology. She has published 3 papers in international journals and 3 papers in reputed National Journals. Also, she has presented 2 papers in International Conference and one paper in the National Conference. Her research interests include Image Processing, Mobile Ad hoc Networks and Security.

P. Ranjetha received B.Tech degree in Information Technology from K.S. Rangasamy College of Technology, affiliated to Anna Anna University, Chennai, Tamil Nadu, India in 2014 and M.Tech degree in Information Technology from K.S. Rangasamy College of Technology, affiliated to Anna University, Chennai, Tamil Nadu, India in 2016. Now she is working as Assistant Professor in Information Technology department at K.S. Rangasamy College of Technology. She has presented two papers in National level technical symposium. Her Research interests include Mining Medical data, Opinion Mining and Web mining.