Machines and Algorithms
http://www.knovell.org/mna

Review Article

It's all about Mobile Databases - Survey

Sheikh Abdul Hannan [1], *

[1] Virtual University, Lahore, 54000, Pakistan
*Corresponding Author: Sheikh Abdul Hannan. Email: ms090400008@vu.edu.pk
Received: 02 November 2022; Revised: 02 December 2022; Accepted: 30 January 2023; Published: 6 March 2023
AID: 002-01-000016

Abstract: As the use of mobile databases is increasing and demanding now a days, so there is need of efficient mobile DBMS. This paper presents useful work done in mobile database systems. Introduction of Mobile DBMS is discussed in this paper. Important characteristics and limitations of any Ubiquitous DMBS have been explored. This survey focuses on various types of mobile DBMS architectures discussed by different authors in different literature. It also addresses different mobile transaction management techniques and models. There is an overview of different concurrency control techniques included, so that data can be managed concurrently. There are different challenges and future concerns in this field and those should be addressed properly.

Keywords: Mobile DBMS; Mobile Transaction; Concurrency Control; Mobile Database Architecture; Ubiquitous DBMS; Mobile Transaction Models;

1. Introduction

Mobile technology and wireless networks are progressing rapidly and a lot of databases driven mobile applications are in the market. People used these mobile applications for ease of their everyday life. Like a manager can manage his resources and information to accomplish the tasks while travelling. A sales representative can update his/her recent sale before leaving the client's place. People can maintain their bank accounts, businesses, social activities and personal data with the help of these applications. Now several companies are developing conventional business-oriented applications as mobile systems to use on handheld devices.

Due to this vast advancement and usage of mobile applications, there is a great need of effective resourceful mobile databases. These databases should handle huge market data and meet the market necessities and challenges efficiently. In most applications, database systems are being used behind; even in the applications like web browsers, antivirus software etc. These mobile database systems are not being used only in these types of applications, these can be useful in small operative gadgets or ubiquitous devices such as smart watches, smart phones, tablets, GPS navigators, sensors, MP3 players etc. As these databases are for devices with mobility options, so these can be termed as Ubiquitous Database Management Systems (UDBMS) [1].

These mobile databases are different in the structure, transaction management, query optimization and other aspects as compared to immobile or traditional databases. The challenges and considerations are also dissimilar and complex for mobile databases. The mobile devices face disconnection frequently while roaming between different cells. The bandwidth of wireless network channels, is also low. On the other hand, these types of devices have limited computation power and small storages. The energy resource like

https://doi.org/10.66108/mna.v2i1.34

battery of these devices is also limited; and it may cause for long disconnections [2-3]. So, there is a need of light weight and strong DBMS, which should meet these challenges [4-5].

The rest of the paper is organized as follows. The next section provides the background and related work information. The section after that discussed the architecture of mobile database. Transaction management and Concurrency control techniques will be discussed in the same section. In the next section, some Mobile DBMSs in the market will be discussed. The subsequent section, summarizes and concludes the paper along with future research horizons.

2. Related work

As per study of different work done already, referenced in the end, Badrinath has presented a new architecture for mobile database which addressed the issues like disconnection, concurrency control and replication management etc [6], [7]. An object-oriented design architecture which supports conflict-free and context-aware data and transaction management, is proposed by Steffen [8]. A way to transform a conjunctive projection-selection query into a unified form, is described by Lee [9]. This transformation proposed two cache partitioning schemes to reduce the granularity of a caching unit and described the semantics of a cache fragment. Ziyad adapted Encryption and Decryption algorithms between base station (BS) and mobile host (MH) for secured data transmission [11]. He worked on security management for mobile database transaction. The same work is further carried and explored by Priya [10].

Sheini worked on the algorithm to predict data in mobile databases [12], [17]. As mobile systems can be in disconnected mode while a transaction, so he proposed Memetic algorithm to predict till to connect to the next mobile network. Yendluri enhanced the conventional optimistic concurrency control (OCC) algorithm through the use of invalidation reports [13]. With the help of these invalidation reports, conflicting transactions can be identified timely and terminated before reaching the validation phase.

Imam proposed Heterogeneous Mobile Database Synchronization Model (HMDSM) for data synchronization between Mobile Clients and Server-side database [14]. In this approach, a listener is added to listen any operation performed on either side. IDBSync (Improved Database Synchronization) was produced by Kottursamy [16], in which the control plane of Software Defined Networking (SDN), synchronizes information across the data planes of mobile hosts and server. It uses BaSyM (Batch Level Synchronization Methodology) table to synchronise different data planes of mobile clients.

Malladi applied multiple query optimization algorithms on mobile databases and analyse the comparative results for optimal solution [15]. Issues and research direction have been addressed by Bernard [2].

2.1. Mobile Database Management System

The applications for ubiquitous devices required totally different ways of data access and management as compared to the traditional DBMSs.

2.1.1. Characteristics of Mobile DBMS

Different characteristics of Mobile DBMS has been discussed before. The authors and researchers have reviewed the main characteristics and stated as the basic requirements of any Mobile DBMS's development. Based on these existing surveys [1], [18-19], the important requirements for ubiquitous DBMS are as follows.

- Streamlined DBMS Integration - These databases functionalities should be integrated with in the application infrastructure, so that there will be no need for further configuration or administration. These databases must be incorporated as DLLs in applications and implemented as a single stand-alone DBMS capable of handling many transactions and applications.
- Mobile Database Footprint - As these databases will be developed for ubiquitous devices with limited storage capacity and does not need continuous connection with the server, so these devices download the required small chunk of information from server and uploads the manipulated data

on server. This small chunk of data downloaded and maintained by mobile database, is called footprint of database [20]. It is important to minimize the size of footprint, so the efficient and simple query execution can be achieved.

- Universal Mobile Compatibility - There is a variety of mobile devices with memory, disk, processor and operating system limitations; so, the mobile DBMS must be able to run on all devices and can be synchronized with back-end systems efficiently.
- Componentized DBMS - As these DBMSs must contain support for small footprints, thus, it is preferable to offer only the functionality necessary by the application. This can be achieved with the help of componentization of DBMS. For example, Simple apps do not require a query processor since they just require record-oriented access. So, these types of DBMSs should support component-based development.
- Self-Sufficient Database System - This DBMS in not visible to end users and usually there is no database administrator (DBA) to manage the database operations. This DBMS should be self-manageable, so that it can perform maintenance operations (backups, recovery, indexing, tuning etc.) by itself or through application. This DBMS should be installed automatically once the application is installed and there will be no need to install this DBMS explicitly or separately.
- In-Memory DBMS Techniques - These DBMSs should serve the applications which need high performance on the data, which will be small enough in the main memory. There should be in-memory DBMSs techniques for optimal query processing and indexing for the small main memory data that may never get persisted.
- Portable Mobile Databases - The mobile applications have to be deployed quickly. When you install these applications, the related database should be installed automatically. So, the mobile database should be highly portable and should be ported with the application as a single file. It means copying the database file completes the migration process of application and there should be no need to install database separately.
- Code-Free DBMS - These DBMSs should be portable and this portability may be act as a carrier of virus or other risks, so there should be no executable code stored in the database. These DBMSs should prevent to store any code in the database.
- Efficient Data Synchronization - These DBMS should facilitate efficient synchronization between mobile hosts and back-end server databases. The data may be retrieved from the back-end database and synced back to the server database after alteration.
- Centralized DBMS Control - As mobile DBMS must be self-managed; so, it should be managed remotely also. The centralized remote management is necessary to configure and manage the mobile databases according to company standards.
- Adaptable Query Systems - There should be facility for custom programming interfaces. As there is a variety of query languages and data models, such as relational, XML, Object-oriented, Streams; it is therefore necessary to modify and componentize these DBMSs in order to provide domain-specific programming interfaces and query languages.

2.1.2. Architecture of Mobile Database

Badrinath consider the centralized database and extended it as Mobile Database, such that mobile devices (hoard clients) can also use this database [6-7]. Hoard clients are the ubiquitous devices which can download (hoard) data and reintegrate the manipulated data back to the server. According to him, there is one centralized database, which can be accessible by traditional wired clients and hoard clients simultaneously.

Traditional clients should be connected always to the server to perform any single database related task. If these clients get disconnection, these cannot perform any operation on the data. On the other hand, the hoard clients get data (footprint) from the centralized server as local database. These clients hoard the small

required fragments only, not the entire data. These clients can be disconnected from the server and perform any operation on local hoarded database while disconnected. Once these clients are connected, these behave like traditional clients and have full access over centralized database. These clients reintegrate the data with centralized server; such as update the manipulated data on the server or hoard updates from the server to refresh its local database etc.

Kumar presented a reference architecture for mobile database that a fixed traditional host should be interconnected through a wired network and mobile hosts should be entertained by a mobile switching centre (MSC) [21-22]. Selvarani further categorized the types of mobile database architecture; like Client/Server architecture, Server/Agent/Client architecture and Peer to Peer architecture.

- Client/Server Architecture

There is one centralized database server. The fixed clients and mobile clients communicate with this server for database operations [23-24]. Yang presented Publisher/Subscription framework in a client/server fashion; where the Publisher objects acts as server and Subscriber objects treated as clients [25]. Zhang also described the 3-tier architecture as client/server architecture; this architecture addressed the synchronization or replication of data between server and mobile hosts [26].

As mobile hosts can hoard data in local database, manipulate the data on local site and update the server database. So, it may lead 3 more architectures. If the server is responsible for all computations and clients with limited resources, need to run operations on the server, then this is Thin Client Architecture. If the clients have complete access and can conduct all server functions while disconnected, then it is Full Client Architecture. Once the responsibilities are divided between clients and server, this is Flexible Client Server Architecture.

- Server/Agent/Client Architecture

An agent is added in this design to execute specified activities. It is a computer programmer who can work on either the server or client side. So, there are three layers in this mobile database architecture; Mobile Terminal Layer, Mobile Agent Layer and Server Layer [27].

- Peer to Peer architecture

Clients in this architecture can communicate with one another. These customers can exchange the information; due to this higher degree of availability of data can be achieved. The consistency should be considered and may be compromised if nodes (clients) update any data.

2.1.3. Transaction Management

Mobile Transaction is a set of multiple operations for a single unit of work in a mobile environment. The main characteristics of the mobile transaction are provided as follows [21-22], [28]:

- As there are frequent disconnection and mobility of data and users exist in this mode, so the mobile transactions should be long-lived.
- A mobile transaction can be divided into several sub-operations and can be run in the mobile host or other hosts on the mobile service station (MSS). This shares its state and partial results with other sub-transactions, so the disconnection and mobility issues can be overcome.
- The computations and communication standards, needed by a mobile transaction, should be compatible with mobile service station (MSS).
- The state of transaction, states of accessed data objects and location information should also move along with mobile host during the mobility of mobile host from one cell to another.
- The mobile transaction should manage the data concurrency and recovery. It should handle disconnection and mutual consistency of replicated objects.

A mobile transaction model is quite complex because of the mobile environment limitations; such as low bandwidth, frequent disconnections and limited resources of mobile hosts. Mobile host can be operated

in fully connected mode, disconnected mode or partially connected mode. Another mode is doze mode, in which it saves energy to minimize the connection cost.

Following are the main models for mobile transactions have been proposed.

- Pre-Write Transaction Model - This model supports once the mobile host initiated the transaction and executes it. It means mobile host gets full autonomy. This concept has offered two data versions, pre-write and write, to boost data availability. [29].
- High Commit Mobile (HICOMO) Transaction Model - HICOMO model is also used to handle mobile transactions in fully autonomous mobile hosts. Aggregate table is used in this model [30].
- Moflex Transaction Model - This model is used in the scenario when mobile transaction is initiated by mobile host but fully executed by data server or fixed host. It supports sub transactions or queries which are location dependant [31].
- Pre-Serialization Transaction Model - This model also suitable for the scenario when mobile transaction initiated by mobile host but executed by data server or fixed host. This model allows local site transactions to commit independently of the global transaction. Local resources can be released in a timely fashion [32].
- Multiversion Transaction Model - In this model, mobile transaction may access stale data at fixed host or database server. This model is suitable once the mobile transaction is executed between mobile hosts and database server or fixed host [33].
- Twin Transaction Model - This model satisfies the both modes of operation, connected and disconnected. Resynchronization mechanism makes sure the consistent state of data [34].
- P system-based Mobile Transaction Model (PMTM) - This model is proposed by Zheng [35]. There are two transitions rules (Membrane Rule and Object Rule) are introduced. Membrane rule describes the structure of membrane and object rule describes the transitions.
- Adaptable Mobile Transaction Model - Alvarado presented this approach in which the transaction characteristics of atomicity and isolation are reduced yet conflict serializability is preserved. [36]. It improves the commit possibilities cost effectively.
- Shadow Paging Transaction Model - Hegazy discussed the Mobile-Shadow technique for mobile transactions [37]. This classifies the actions to be performed during validation phase of a transaction.
- Surrogate Object Based Mobile Transaction Model - Ravimaran proposed a methodology to facilitate data caching at surrogate objects in order to retrieve data more quickly and conduct database operations more efficiently [38].
- Connection Fault-Tolerant Model - This architecture decreases resource blocking time at the Server or fixed host. It allows for quick connection recovery and increases the number of committed mobile transactions. Dimovski proposes this concept [39].
- Kangaroo Transaction Model - In this concept, a new layer (data access agent) is added to all base stations on top of the current global transaction manager. It addresses the mobility issue during mobile transactions [40].
- Clustering Transaction Model - In this model, the mobile transaction is divided into two types, strict transaction and week transaction. Strict transactions are performed in connected mode only whereas week transactions execute operations during disconnected mode and synchronize the database once connected [40].
- Two-Tier Transaction Model - Lazy replication mechanism has been used in this model. Tentative transactions are performed in disconnected mode and on reconnect these tentative transactions are converted to base transactions (transactions performed on the master version of data).

2.1.4. Concurrency Control

Concurrency control can be achieved through implementation of two phases locking scheme. This locks all the data and transaction, for a specific time period. Another technique is pessimistic concurrency control, which assumes that the data utilized by one transaction may be needed by others and hence locks the data. Once the transaction committed then the lock should be released.

Optimistic concurrency control technique does not lock the data and same data can be accessed by different transactions concurrently. On commit time conflict should be checked and resolved. Another strategy is to use as low an isolation lever as feasible in order to archive good performance and availability. Shot Duration Locks are also implemented to increase the concurrency among transactions. These locks will be released as soon as possible.

2.2. Mobile DBMSs in market

There are many mobile DBMSs are already in the mobile application's market serving the mass. Some of the famous DBMSs are following:

- SQLite – an open-source embedded lightweight database engine suitable for embedded applications. Size of its footprint is just 500KB. It does not need any configuration or any installation. It also restores automatically after system collapse as it does not need any DBA also. It is plate-form independent and supporting multiple tables and indexes, transactions and ACID property, views, triggers etc.
- SQL Anywhere – an embedded database system and can be deployed with the application installation. It has less than 275KB footprint. It can be called through a simple API. Its transaction log and data are stored in separate files. It provides 128-bit encryption security for data objects.
- IBM DB2 EVERYPLACE – is consisting of shared libraries and used as APIs. Its footprint is 350KB. Compiler and Data Manager Service is used to abstract, parse, optimize and execute query plans.
- ORACLE DATABASE LITE – includes small footprint, 2MB, and used for small business environments. It has data synchronization capabilities and can be installed via self-extracting setup.exe files. It is intended for use in embedded devices and has a compact footprint. There is no involvement of DBA.
- SQL SERVER COMPACT – is a redistribute compact relational database for embedded applications. OLE DB allows applications to connect to the database. It provides ACID properties and 128-bit encryption.

Table 1: Summary of the Key Finding

Aspect	Findings
Architectures & Approaches	- Badrinath's architecture addresses disconnection, concurrency control, and replication management. - Steffen's object-oriented design supports conflict-free, context-aware data, and transaction management. - Lee's query transformation and cache partitioning. - Ziyad and Priya's work on encryption for secure data transmission. - Sheini's algorithm using Memetic algorithm for predicting data in disconnected mobile transactions. - Yendluri's enhancement of optimistic concurrency control. - Imam's Heterogeneous Mobile Database Synchronization Model (HMDSM). - Kottursamy's IDBSync with SDN for synchronization. - Malladi's multiple query optimization algorithms. - Bernard's addressing of issues and research directions.
Characteristics of	- Integration with application infrastructure. - Minimal footprint (small chunk of information downloaded and maintained). - Support for various devices. -

Mobile DBMS	Self-management for maintenance operations. - Component-based development for functionality. - In-memory DBMS techniques for optimal query processing. - Portability with quick deployment. - No executable code stored to prevent risks. - Efficient synchronization between mobile hosts and back-end server databases. - Remote management for configuration according to company standards. - Custom programming interfaces for various query languages and data models.
Architecture of Mobile Database	- Different architectures: Client/Server, Server/Agent/Client, Peer-to-Peer. - Centralized database with hoard clients (Badrinath). - Reference architecture with fixed traditional hosts and mobile hosts (Kumar). - Categorized types: Client/Server, Server/Agent/Client, Peer-to-Peer (Selvarani).
Transaction Management	- Various mobile transaction models with different characteristics and complexities: Pre-Write, HICOMO, Moflex, Pre-Serialization, Multiversion, Twin, P system-based, Adaptable, Shadow Paging, Surrogate Object-Based, Connection Fault-Tolerant, Kangaroo, Clustering, Two-Tier. - Characteristics include long-lived transactions, compatibility with mobile service stations, handling of disconnection and mobility issues, and management of data concurrency and recovery.
Concurrency Control	- Two-phase locking scheme. - Pessimistic concurrency control. - Optimistic concurrency control. - Low isolation levels for performance and availability. - Short Duration Locks for increased concurrency.
Mobile DBMSs in the Market	- SQLite: open-source, lightweight, embedded, platform-independent. - SQL Anywhere: embedded, less than 275KB footprint, 128-bit encryption. - IBM DB2 EVERYPLACE: shared libraries, APIs, 350KB footprint. - ORACLE DATABASE LITE: small footprint (2MB), embedded, intended for embedded devices. - SQL SERVER COMPACT: redistributable compact relational database, supports ACID properties, and 128-bit encryption.

Summary

The referenced works present various approaches and solutions in the domain of mobile database management systems (DBMS). Badrinath introduces a novel architecture addressing disconnection, concurrency control, and replication management issues in mobile databases. Steffen proposes an object-oriented design architecture supporting conflict-free and context-aware data and transaction management. Lee focuses on transforming conjunctive projection-selection queries into a unified form, presenting cache partitioning schemes to optimize caching unit granularity. Ziyad and Priya contribute to security management in mobile database transactions through encryption and decryption algorithms. Sheini develops an algorithm using a Memetic approach to predict data in mobile databases, considering disconnected modes. Yendluri enhances optimistic concurrency control using invalidation reports to identify and terminate conflicting transactions timely. Imam introduces the Heterogeneous Mobile Database Synchronization Model for efficient data synchronization between mobile clients and server-side databases. Kottursamy's IDBSync employs Software Defined Networking (SDN) control planes for information synchronization across mobile hosts and servers. Malladi explores multiple query optimization algorithms for mobile databases, analysing comparative results for optimal solutions. Bernard addresses issues and research directions in the field. The subsequent section details the characteristics, architecture, transaction management models, concurrency control strategies, and existing mobile DBMSs in the market. Key findings include the importance of self-manageability, efficient synchronization, and footprint minimization for ubiquitous DBMSs, as well as diverse transaction management models to address the challenges posed by the mobile environment. Additionally, various architectures, such as Client/Server, Server/Agent/Client, and Peer to Peer, cater to different communication and data exchange scenarios in

mobile databases. The market section lists several mobile DBMSs like SQLite, SQL Anywhere, IBM DB2 EVERYPLACE, ORACLE DATABASE LITE, and SQL SERVER COMPACT, each with unique features catering to specific needs in the mobile application market.

3. Conclusions

In this paper, introduction of mobile databases has been provided. Important characteristics of a mobile DBMS are discussed in detail which leads to a better design of M-DMBS. Architecture of M-DBMSs is different than traditional DBMS. As M-DBMSs can work in disconnected mode also, so it should hoard data locally in form of footprints. After manipulation it should reintegrate the data on server.

Transaction management in mobile environment is also a challenge as mobile devices can be disconnected from the network, moved from once cell to another. Different models to handle the mobile transaction have been discussed and provided the techniques for concurrency control.

There are still many future concerns and research challenges should be addressed. Efficient Data replication and synchronization techniques are needed. Transaction management and concurrency control, is a hot topic for future research. There is a need of efficient light weight DBMS with small footprints. How to get data confidentiality, privacy and security over the wireless network.

References

[1] Whang, Kyu-Young, Il-Yeol Song, Taek-Yoon Kim, and Ki-Hoon Lee. "The ubiquitous DBMS." *ACM SIGMOD Record* 38, no. 4 (2010): 14-22.

[2] Bernard, Guy, Jalel Ben-Othman, Luc Bouganim, Gérôme Canals, Sophie Chabridon, Bruno Defude, Jean Ferrié et al. "Mobile databases: a selection of open issues and research directions." *ACM Sigmod Record* 33, no. 2 (2004): 78-83.

[3] Barbará, Daniel. "Mobile computing and databases-a survey." *IEEE transactions on Knowledge and Data Engineering* 11, no. 1 (1999): 108-117.

[4] Araujo, Mozart, Clauirton A. Siebra, Fabio QB da Silva, and Andre LM Santos. "A lightweight DBMS solution for mobile devices." In *2009 IEEE International Conference on Communications Technology and Applications*, pp. 946-951. IEEE, 2009.

[5] Tavakkoli, F., A. Andalib, A. Shahbahrami, and R. Ebrahimi Atani. "A Comparison of Lightweight Databases in Mobile Systems." (2011).

[6] Badrinath, B. R., and Shirish Hemant Phatak. "A Database Architecture for Handling Mobile Clients1." (1998).

[7] Badrinath, B. R., and Shirish Phatak. *An architecture for mobile databases*. Rutgers University, 1998.

[8] Vaupel, Steffen, Damian Wlochowitz, and Gabriele Taentzer. "A generic architecture supporting context-aware data and transaction management for mobile applications." In *Proceedings of the International Conference on Mobile Software Engineering and Systems*, pp. 111-122. 2016.

[9] Lee, Ken CK, Hong Va Leong, and Antonio Si. "Semantic query caching in a mobile environment." *ACM SIGMOBILE Mobile Computing and Communications Review* 3, no. 2 (1999): 28-36.

[10] Priya, A. "Security Management System for Mobile Database Transaction Model using Encryption and Decryption Algorithm." *International Research Journal of Engineering and Technology* 2, no. 05 (2015).

[11] Abdul-Mehdi, Ziyad T., and Ramlan Mahmod. "Security Management Model for Mobile Databases Transaction Management." *2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications*. IEEE, 2008.

[12] Sheini, Leila Alaei, Khomein Branch Engineer, Hamid Paygozarh, and Mohammad Khalily Dermany. "New Genetic Algorithm to Predict Data In Mobile Databases."

[13] Yendluri, Anand, Wen-Chi Hou, and Chih-Fang Wang. "Improving concurrency control in mobile databases." In *Database Systems for Advanced Applications: 9th International Conference, DASFAA 2004, Jeju Island, Korea, March 17-19, 2003. Proceedings, 9*, pp. 642-655. Springer Berlin Heidelberg, 2004.

[14] Imam, Abdullahi Abubakar. "Data Synchronization Model for Heterogeneous Mobile Device Databases And Server-Side Database." Phd Diss., Universiti Teknologi Petronas, 2015.

[15] Malladi, Rajeswari, and Karen C. Davis. "Applying multiple query optimization in mobile databases." In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*, pp. 10-pp. IEEE, 2003.

[16] Kottursamy, Kottilingam, Gunasekaran Raja, Jayashree Padmanabhan, and Vaishnavi Srinivasan. "An improved database synchronization mechanism for mobile data using software-defined networking control." *Computers & Electrical Engineering* 57 (2017): 93-103.

[17] Sheini, Leila Alaei, Khomein Branch Engineer, Hamid Paygozarh, and Mohammad Khalily Dermany. "A Multi-Objective Optimization Algorithm to Predict Information in Mobile Databases."

[18] Nori, Anil. "Mobile and embedded databases." In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of data*, pp. 1175-1177. 2007.

[19] Wankhade, Nitin V., and S. P. Deshpande. "A Review on Databases for Mobile Devices." *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)* (2015): 276.

[20] Birari, Nilesh. "Small footprint of Database on tablet." PhD diss., Indian Institute of Technology, Bombay, 2013.

[21] Kumar, Vijay. *Mobile database systems*. John Wiley & Sons, 2006.

[22] Selvarani, D. Roselin, and T. N. Ravi. "A survey on data and transaction management in mobile databases." *arXiv preprint arXiv:1211.5418* (2012).

[23] Xia, Yanli, and Abdelsalam Helal. "A dynamic data/currency protocol for mobile database design and reconfiguration." In *Proceedings of the 2003 ACM symposium on Applied computing*, pp. 550-556. 2003.

[24] Chan, Darin, and John F. Roddick. "Summarisation for mobile databases." *Journal of Research and Practice in Information Technology* 37, no. 3 (2005): 267-284.

[25] Li, Yang, Xuejie Zhang, and Yun Gao. "Object-Oriented Data Synchronization for Mobile Database over Mobile Ad-hoc Networks." In *2008 International Symposium on Information Science and Engineering*, vol. 2, pp. 133-138. IEEE, 2008.

[26] Zhang, Yun, Mu Zhang, and Fuling Bian. "A tiered replication model in embedded database based mobile geospatial information service." In *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*, pp. 1-4. IEEE, 2008.

[27] Li, Jing, and Jianhua Wang. "A new architecture model of mobile database based on agent." In *2009 First International Workshop on Database Technology and Applications*, pp. 341-344. IEEE, 2009.

[28] Mobarek, Adil, Siddig Abdelrhman, Areege Abdel-Mutal, Sara Adam, Nawal Elbadri, and Tarig Mohammed Ahmed. "Transaction processing, techniques in mobile database: An overview." *Int. J. Comput. Sci. Appl* 5 (2015): 1-10.

[29] Madria, Sanjay Kumar, and Bharat Bhargava. "A transaction model to improve data availability in mobile computing." *Distributed and Parallel Databases* 10 (2001): 127-160.

[30] Lee, Minsoo, and Sumi Helal. "Hicomo: High commit mobile transactions." *Distributed and Parallel Databases* 11 (2002): 73-92.

[31] Ku, Kyong-I., and Yoo-Sung Kim. "Moflex transaction model for mobile heterogeneous multidatabase systems." In *Proceedings Tenth International Workshop on Research Issues in Data Engineering. RIDE 2000*, pp. 39-45. IEEE, 2000.

[32] Dirckze, Ravi A., and Le Gruenwald. "A pre-serialization transaction management technique for mobile multidatabases." *Mobile Networks and Applications* 5 (2000): 311-321.

[33] Lam, Kam-Yiu, Guo Hui Li, and Tei-Wei Kuo. "A multi-version data model for executing real-time transactions in a mobile environment." In *Proceedings of the 2nd ACM international workshop on Data engineering for wireless and mobile access*, pp. 90-97. 2001.

[34] Cuce, Simon, Arkady Zaslavsky, Bing Hu, and Jignesh Rambhia. "Maintaining consistency of twin transaction model using mobility-enabled distributed file system environment." In *Proceedings. 13th International Workshop on Database and Expert Systems Applications*, pp. 752-756. IEEE, 2002.

[35] Zheng, Baihua, Wang-Chien Lee, and Dik Lun Lee. "On semantic caching and query scheduling for mobile nearest-neighbor search." *Wireless Networks* 10 (2004): 653-664.

[36] Serrano-Alvarado, P., Roncancio, C., Adiba, M., & Labbé, C. (2005). An Adaptable Mobile Transaction Model. *Computer Systems Science and Engineering (IJCSSE)*, *20*(3), ISSN-0267.

[37] Hegazy, Osman Mohammed, Ali Hamed El Bastawissy, and Romani Farid Ibrahim. "Handling Mobile Transactions with Disconnections Using a Mobile-Shadow." *Cairo University, Egypt* (2008).

[38] Ravimaran, S., and MA Maluk Mohamed. "An improved kangaroo transaction model using surrogate objects for distributed mobile system." In *Proceedings of the 10th ACM International Workshop on Data Engineering for Wireless and Mobile Access*, pp. 42-49. 2011.

[39] Dimovski, Tome, and Pece Mitrevski. "Connection Fault-Tolerant Model for distributed transaction processing in mobile computing environment." In *Proceedings of the ITI 2011, 33rd International Conference on Information Technology Interfaces*, pp. 145-150. IEEE, 2011.

[40] Serrano-Alvarado, Patricia, Claudia Roncancio, and Michel Adiba. "A survey of mobile transactions." *Distributed and Parallel databases* 16, no. 2 (2004): 193-230.

Machines and Algorithms
http://www.knovell.org/mna

Research Article

Service Discovery Framework for Fog Computing

Rizwan Amin[1, *], Asif Raza[1] and Tanzeela Kausar[2]

[1]Department of Computer Science, Bahauddin Zakariya University, Multan,60000, Pakistan
[2]Institute of Computer Science & Information Technology, The Women University, Multan,60000, Pakistan
[*]Corresponding Author: Rizwan Amin. Email: muhammadrizwanbrw@gmail.com

Received: 09 November 2022; Revised: 07 December 2022; Accepted: 03 February 2023; Published: 6 March 2023
AID: 002-01-000017

Abstract: Fog computing has experienced significant growth, enabling access to cloud-based services from nearby fog nodes instead of relying on centralized cloud systems. This decentralization allows for the distribution of services across various systems, providing customers with increased proximity and efficiency in accessing the resources they need. In addition to the shared skills of cloud computing, fog computing gives additional features that facilitate mobility, low latency, and real-time interactions. Nevertheless, it remains a tough issue to identify distributed fog computing resources. The purpose of this research was to develop a fog computing service discovery framework that makes use of machine learning. Random Forest, Logistic Regression, GridSearchCV, and Support Vector Machine are only a few of the current machine learning algorithms that the suggested framework uses in the service discovery process. Results are evaluated and contrasted based on the suggested learning model's performance. Because of their specific goals, Random Forest and GridSearchCV are our primary ML algorithms, even if there are others that show superior accuracy.

Keywords: IOT; Cloud Computing; Fog Computing; Service Discovery; Target access best node;

1. Introduction

The Internet of Things (IoT) refers to a community of physical devices which are interconnected with other internet-linked devices and structures that use sensors, software, and technology. The IoT revolution has the capability to create new market possibilities and enterprise fashions in areas such as safety, privacy, and technological interoperability. The extensive adoption of IoT devices is expected to impact various facets of our daily lives. These devices consist of secondary customers and innovative objects like power supplies, home automation, and internet access devices. It is expected that the global economic system of IoT will grow from $3.9 trillion to $11.1 trillion by 2025 [1].

Because the IoT continues to grow, the idea of edge devices or fog computing devices has emerged. Fog computing is one of the cloud models which have received popularity due to the growing interest in cloud computing. Fog computing allows the advent of low-latency network connections in computing devices, analytics, and nodes [2]. Further, network devices, cloud servers, and fog computing nodes can offer diverse services such as region detection, aspect transmission, facilitating mobility, low latency, and real-time interactions.

However, fog computing resources are allotted throughout a multi-layer network and are managed by multiple owners, making the resource discovery problem even more complex. Decentralized resource

https://doi.org/10.66108/mna.v2i1.38

discovery is critical to deal with this problem [4]. Several systems and toolkits for fog computing have been proposed, however researchers still face problems identifying the precise ambiguous nodes in a service. Many researchers discuss the services discovery with quick access to nodes, reduced communication latency, improved performance and security during information transformation, and authentication [5]. However, most of the models do not focus of the accuracy and most beneficial serving nodes.

To address these challenges, we endorse a service discovery infrastructure for fog computing, using logistic regression algorithms, support vector machines, GridSearchCV, and random forest techniques.

The rest of the sections of the paper are presented as follows Section II gives you background study, Section III a literature overview, Section IV the proposed framework the results and experiments of the version. Ultimately, Section V contain conclusion.

2. Background Study

The Internet of Things (IoT) is a network of physical things that are integrated in hardware and software that can connect to and exchange data with other devices and computers on the internet. Because of low-cost computing, cloud computing, big data analytics, and mobile technologies, all appliances, from simple household items to cutting-edge industrial machines, can share and retain data with minimum human participation. Sensors that are both affordable and reliable have made IoT technology more accessible to a broader variety of manufacturers.

The Internet of Things (IoT) is a growing technology [7]. Millions of petabytes of data are collected every day by billions of networked devices throughout the world. This data might contain vital information about home safety, entertainment needs, water saving, and decreased fuel emissions. The Internet of Things has been introduced to us through smartphones, live TV services, smart TVs, and other gadgets.

2.1. Computing in the Fog

Fog computing is a sort of decentralized computing in which processing resources are shared between data sources and a cloud or other data center. Fog computing is a way of managing user requests on edge networks. Routers, gateways, bridges, and chassis are examples of fog-filled network-related equipment. These solutions, when implemented by specialists, will be capable of performing both computer and networking operations. Though their capacity is limited in comparison to cloud waitrons, their geological size and regionalized geography enable them to deliver trustworthy services across a wide area. The term "fog" refers to a location that is physically closer to a user's computer system than the cloud [10].

3. Review of Literature

A novel computing architecture known as fog computing increases the processing capabilities of cloud services. There are resources in the form of fog nodes along the beach [1]. For location, fog provides efficient data processing and storage, decreased latency, and network bandwidth savings. One of the criteria for attaining these goals is the identification of optimal fog using the edge tool. Sign one before deciding on the best contract. There are numerous approaches to achieve the needs of these objectives. The goal of this study is to provide a comprehensive assessment of the country. Explore the fog node area to locate and select art. To do this, we create another first. Certain requirements must be followed in order to establish the greatest fog retention. So, we define and compare this to the literature from 2014 to 2021, as well as the specific identifying criteria. There is still a significant quantity of fog. Based on the investigation, we propose gaps and unsolved issues in the available literature. The largest node was chosen and utilized to detect fog.

When distributed systems are deployed on fuzzy computing nodes, microservice architecture becomes an important and difficult challenge. In cloud contexts, service discovery algorithms are widely deployed in computer networks and clusters [3]. Using a service discovery technique, the service container writes its own data in a customer-centric manner. When a request comes in, the services proxies utilize a lookup table

to route it to an available server. Clients in fog computing, on the other hand, are no longer restricted to cloud resources, but can instead request resources from local fog nodes.

Smart buildings are comprised of real-time information systems comprised of a variety of sensors and equipment linked by communication networks [4]. On the other side, its complexity develops. The implementation of Internet of Things (IoT) models has the potential to play a greater role in building energy management by significantly increasing its capacity to acquire, analyze, and display data as knowledge. This complex and growing data requires sophisticated processing resources, which cloud computing models may provide. Although technology is widely utilized, it has problems such as insufficient latency, traffic congestion, and a lack of support for localization. As a result, fog has emerged as a viable solution for offering fluctuating resources at network edges. Intelligent smart building framework for cloud and location, a unified framework for IoT devices placed in buildings to do computing, routing, and simultaneous interaction with apps operating in facilities.

Smart agriculture is a strategy of achieving long-term agricultural and food production that makes use of interactive tools and cutting-edge technology. Smart farming is purely based on the Internet of Things, which relieves farmers and farmers of physical labor and thereby enhances production in every way imaginable. With the present emergence of agricultural trends, the Internet of Things provides enormous benefits such as efficient water consumption, improved resources, and more. The main distinctions are the large benefits and the current occurrence of revolutionary agriculture.

The smart cloud service discovery framework (2020) was developed by Al-Sayed et al. He regards his method as rational, clear, adaptable, and all-encompassing. As we all know, intelligent gadgets make considerable use of artificial intelligence technologies. Standard: Provides all cloud providers with a uniform default identity model for offering their services in a competitive cloud service segment. It requires a flexible limit that allows even non-technical employees to construct text queries and grasp ambiguity and complicated investigations. Because it covers all common functional elements of cloud service providers, this technique gives a graphical interface with little understanding of cloud concepts while addressing the restrictions of all cloud providers. Users may now ask complex questions in their own language. The results show a low error rate, excellent agreement, and the utilization of conventional detection methods.

To address the limitations of previous solutions, Sharma et al. [23] created a blockchain-based cross-cloud discovery platform for non-federated intermediate models. One of the goals is to enhance the design of the P2P structure combination so that CSPs may be formed almost densely. It also implies that using a blockchain for inter-cloud service discovery will eliminate the need for a right-hand third entity or intermediary among participating cloud service providers, with scalable service detection being the primary benefit of the proposed method. This strategy, however, is limited by delayed reflexes.

Finally, Ka et al. [24] propose the SIM SIM service discovery to improve service discovery performance for P2P mobile cloud applications. SimSim uses the same 2D grayscale environment and segmentation to show services. Single-hop routing and thorough environmental research on the effectiveness of internal control in difficult situations. In terms of similar hash maps, the proposed method outperforms previous approaches in terms of clustering. Sorting hashes improves the comparability of layer and stored resources. Because it uses a hierarchical hash aggregation approach to seek installation and discovery methods for services with keywords criterion in the gray space, this technique provides services with comparable geographic proximity. As a result, the SIM outperforms in job scheduling and reaction time. Its weakness, however, is a lack of security.

In addition, in order to explore intricate system behavior, Zahoor et al. [27] proposed UML characteristics for service discovery in an organizational cloud architecture, as well as the synthesis approach necessary for system modeling. Their strategy is to make the system easier to understand and use for consumers, developers, and developers. The goal of this methodology is to give a UML configuration file to describe a discovery method for establishing proxy-based online facilities, as well as the best UML file type that can be used with any discovery strategy. Furthermore, it is not overly complicated, and the mechanism's weakness is its poor response time.

Finally, Abbas et al. [28] provide a cloud discovery service that includes an acceptable job design model and cloud ontology. To some extent, the proposed architecture addresses vendor security issues while also improving cloud portability and interoperability. The multi-agent schema is made up of several smart agents such as Directories Mediator, Detection Mediator, Rating Agent, Consumption Agent, and others. This document describes the methodical approach of creating a reference model. It is used in multi-agent systems to build intelligent agents and provide interoperability. Because it overcomes portability and interoperability [28] difficulties in cloud technology, as well as general supplier security concerns, the findings show that the reference design performs much better in terms of lookup effectiveness, reply time, and implementation time.

A truly decentralized mechanism. Navimipour [26] provided a trust-based strategy to personalizing cloud service discovery and verification techniques. The NuSMV model validator, as well as relevant models, administrative diagrams, historical logic, and the NuSMV model validity for verifying model constructions and elements, are discussed. Because the predicted design and structure of trustworthy HR discovery are established using temporal logic, the approach may efficiently search for effective, available, efficient, and discontinuous services. Clean, fair, and long-lasting. However, this technique disregards quality of service (QoS) and permissions.

To achieve safe service discovery for IoMT, we suggest using SSD blockchain-based fog computing [22]. We suggest a cross-blockchain design comprised of numerous self-governing comparable blockchains, motivated by the premise that a large number of smoke producers may successfully offer clients in various regions. In this situation, each secure concurrent blockchain may demonstrate the reliability of a certain multimedia application at a fair cost. In a recent publication, we developed a secrecy image-based domain query approach for fog-enabled vehicle networks. This research also looks into the improved performance and security of proposed SSDs [25].

4. Proposed Methodology

The main principles and procedures of the suggested approach were covered in this study. All of the methods and strategies utilized in the fog computing service discovery framework must then transit through various phases of the framework layer. The flowchart and algorithm below demonstrate all of the basic steps of the proposed technique.

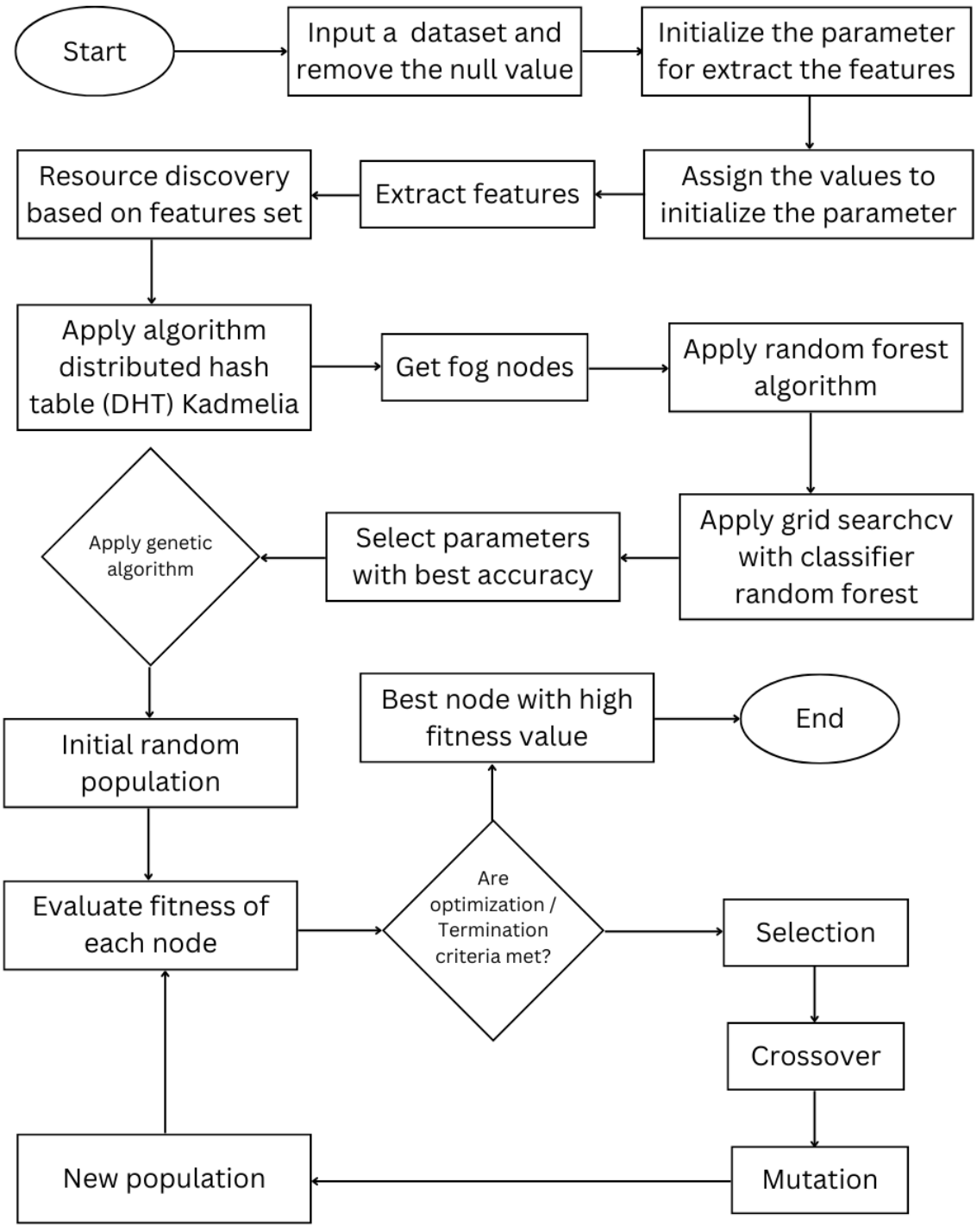

Figure1: Methodology Data Flow Diagram for GA

- Algorithm 1: Service Discovery Framework for Fog Computing

Input: **D** ▷ **D = Dataset**

Output: n is a best matching node

D ← Remove $(Null/NaN, D)$

1. $(X, Y) \leftarrow Extract\ (D)$ ▷ X , Y = Extract the values x,y on dataset
2. $(x\_train, x\_test, y\_train, y\_test) \leftarrow split\ (X, Y)$
3. $S \leftarrow$ Assign (n_estimators, max_features, max_depth, max_samples) ▷ S = Set of Hyper Parameters (n_estimators, max_features,max_depth, max_samples)
4. $S' \leftarrow$Pram_Grid (n_estimators, max_features, max_depth, max_samples)
5. (R, A) ←GridsearchCV (Random Forest, S', S) ▷ R = Best parameter, A = Accuracy
6. n$\epsilon S' \leftarrow$GA(R) ▷ GA = Genetic Algorithm
7. **T**← T Random population ▷ T= Random Population set generated by GA
8. N; ▷ N =No. of Iterations set to predefined values
9. FFV = 5; ▷ FFV = Fitness function value
10. i = 0; ▷ i = set of condition
11. While (FFV! = 0 OR i ≠ N)
12. O←Count (Desired Resources – T[i] ▷ ∀ i ϵ size(T)
    ▷ T[i] = Nodes available resources
13. $O' \leftarrow$ Mutation(O)
14. End
15. Output n ← best of O

- Algorithm 2: Genetic Algorithm

Input: D ▷ **P = Generate Initial Random Population**

Output: n ▷ n = Best Node, high fitness value

1. S ← Param ▷ Param = Set of GA parameters
2. P ← Assign (**S**) ▷ P = Generate Random Papulation
3. F ← P ▷ F = Evaluate Fitness Value
4. While (F == T) ▷ T = Termination Criteria
5. $P' \leftarrow S'$ ▷ S'= Select a new population
6. $C' \leftarrow P'$ ▷ C' = Cross Over
7. $M' \leftarrow C'$ ▷ M' = Mutation
8. n ← M' ▷ n = Generate a new population
9. END
10. **Output** n ← best node with high fitness value

4.1. Remove Empty Values in CSV dataset

The data set was not without flaws. They can end up with values that are incorrect, corrupted, or missing. I have no NULL or NaN values for the project I'm working on. This discussion group will teach you how to construct a simple Python script to check if your dataset contains NULL or NaN values and, if so, how to update the data. He'll offer you alternatives.

4.2. Initialize and Assign the Value of Parameters

At this point, every model we've built has required us to begin with parameters based on a certain distribution. So far, we've examined the initial pattern and how this decision was reached. These things may appear to be unimportant to you. Instead, the initial mode is key for maintaining numerical stability and playing an important role in the learning service discovery architecture.

4.3. Parameter Management

Once we've chosen a structure and defined our hyperparameters, we start a learning loop with the objective of finding parameter values that reduce our loss function. These parameters will be required for future projections following training [15]. We may also need to get parameters in order to reuse them in another framework, save our prototype to disk in order to use it with other programmers, or collect scientific knowledge for testing purposes.

4.3.1. Parameter Access

When a model is defined by a join class, we can access any class by indexing the first model as if it were a slope. In its table, we can easily find the parameters of each coat. As tails, we may test the parameters of the next totally related coating.

4.4. Feature Extraction

Feature extraction is a dimension reduction approach that condenses a huge dataset into a smaller processing set. One advantage of such large datasets is the presence of a high number of variables that require considerable processing resources to execute [15]. Feature extraction is a method of discovering and/or merging parameters into features that efficiently reduces the amount of data that must be addressed while maintaining the correctness and completeness of the real data set.

In order to get to the meat of the data, analysts painstakingly isolated a number of attributes. 'Time,' 'Period Time,' and 'Time since reference,' among others, are temporal components that give a deeper understanding of the time dimension. The frames are subjected to additional analysis utilizing 'Previous Capture Frame' and 'Previous Display Frame,' which allows for a comprehensive evaluation of the data. 'Payload Length' and 'Total Length,' which stand for payload attributes, reveal details on the structure and amount of the material. The 'Date' function, which adds a chronological anchor to the data, completes the picture and makes it easier to analyze.

4.5. Random Forest

Random Forest is a supervised machine learning method. The "forest" they make is made up of deciduous trees that are typically gathered together in "flocks." The main notion of the activation approach is to integrate the learning model in order to optimize the entire effect.

4.6. GridSearchCV

To begin, what is a web search? It is the process of experimenting with totally acceptable hyper - parameters in order to get the optimal values for a certain model. As previously stated, hyperparameter settings have a major influence on model performance. It is important to note that knowing the optimal solution of hyper - parameters in advance is impossible, thus we must test all conceivable values to find the best one.

- Estimator: Pass the model instances for which the hyperparameters should be verified.
- Params_grid: the dictionary object holding the hyperparameters to be tested.
- Scoring: You may just enter a valid string as the evaluation measure you want to employ.
- CV: A specific number of cross-validations must be attempted for each set of hyperparameters.
- Verbose: While fitting the data to GridSearchCV, set it to 1 to receive a detailed report.
- Jobs: If you provide -1 for the number of processes to run in parallel for this task, all available processors will be used.

4.7. Genetic Algorithms

They are commonly employed to provide high-quality solutions to optimization and search problems. Using the genetic method, we obtain an initial random population, which is then utilized to evaluate the fitness value of each node. As a result, two outcomes are possible. If the termination requirements are fulfilled, it shows the best node with a high fitness value; otherwise, the selection and crossover processes are repeated.

4.7.1. Fitness Function

The fitness function assists in determining everyone's fitness. It determines each individual's own fitness score, which influences their chance of reproducing. The more fit you are, the more likely you are to be chosen for fertility.

Fitness function and Node fitness formula, for example, are explained further below.

$$Fitness\ Function\ Value\ =\ Desired\ Resources\ -\ Nodes\ Available\ Resources$$

$$FFV\ =\ 11111 - 111$$

$$FFV\ =\ 11$$

$$FFV\ =\ 2$$

$$Node\ Fitness\ =\ Node\ Available\ Resources\ /\ desired\ Resources$$

$$Node\ Fitness\ =\ 2\ /\ 5$$

$$Node\ Fitness\ =\ 0.40$$

4.7.2. Initial Population

The first stage in the Genetic Algorithm is to populate the population. Our suggested algorithm incorporates this step. Genetic algorithms produce a random population. We analyze the population of the 51 nodes and compute the fitness value for each node.

$$Population\ population\ =\ new\ Population(\ )$$

$$Demo.population.initialize(51)$$

$$\sum_{i=0}^{L} new\ individual(\ )$$

4.7.3. Selection

In this procedure, parents are picked at random from the current population.

$$P(N_{51}) = \left| \frac{P(N_{51})}{\sum_{i=1}^{n\ pop} f(N_o)} \right|$$

4.7.4. Crossover

At this point, the tails of the two parents are exchanged to make new offspring at a randomly determined crossover site and pick the next population among people.

$$int\ crossOverPoint\ =\ rn.nextInt(population.individuals[0].nodeLength)$$

4.7.5. Mutation

In the process of mutation, the resources are replaced with one another. Mutation derives the number from the total resources of the nodes. When the value is 0, it gets replaced with 1, and vice versa.

$$int\ mutationPoint\ =\ rn.nextInt(population.individuals[0].nodeLength)$$

5. Experiments and Results

We receive the results after applying the machine learning model to the given dataset. The Random Forest Algorithms that we used on our Model. However, in order to compare the results, we also tried various different Algorithms on the data set.

Other algorithms provide superior accuracy, but our focus is on Random Forest and GridSearchCV, which are used for certain purposes and methods.

5.1. Accuracy of the Training Model with Train and Test Datasets

First, we split the train and test data sets in the original model. And we allocate 80% of the knowledge set to training and 20% to testing. When we tested logistic regression, we achieved 59% accuracy. Then I tested Support Vector Machine, which had a 60% accuracy rate. Then I looked into Random Forest, which has a 62% accuracy rate. GridSearchCV was tested and found to be 72% accurate, which is higher than any other algorithms.

After Applying the Machine Learning Model on Given Dataset as we get the Results. The Algorithms that we applied it Random Forrest on our Model. But for compare the Results we try to some other Algorithms as well on data set.

The other Algorithms is also giving better accuracy but our focus is on Random Forest and GridSearchCV that applied with specific purpose and Methods.

5.2. Training Model's Accuracy

As evidenced by the accuracy table. Logistic Regression has a 59% accuracy, Support Vector Machine has a 60% accuracy, Random Forest has a 62% accuracy, and GridSearchCV has a greater accuracy than all of them together.

Table 1: Training Model's Accuracy

Models	Accuracy
GridSearchCV	72%
Random Forest	62%
Support Vector Machine	60%
Logistic Regression	59%

5.3. Confusion Matrix

A confusion matrix is a table which is used to describe a classification procedure's presentation. The effectiveness of a cataloging model is shown and brief using a confusion matrix. Values show a confusion matrix with benign nodes accessible and not accessible is given below

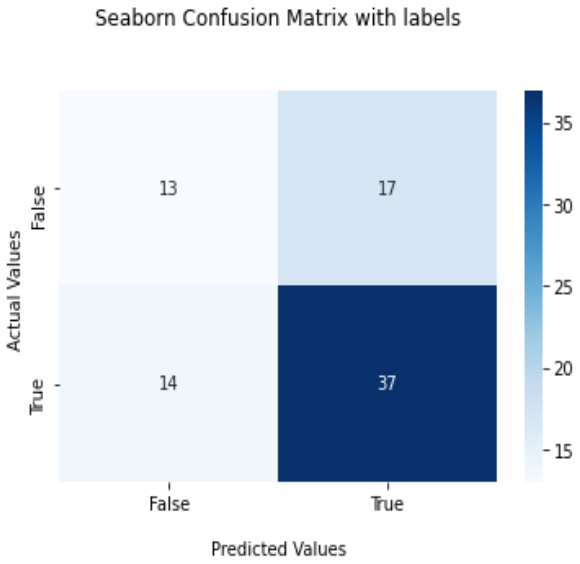

Figure 2: Confusion Matrix

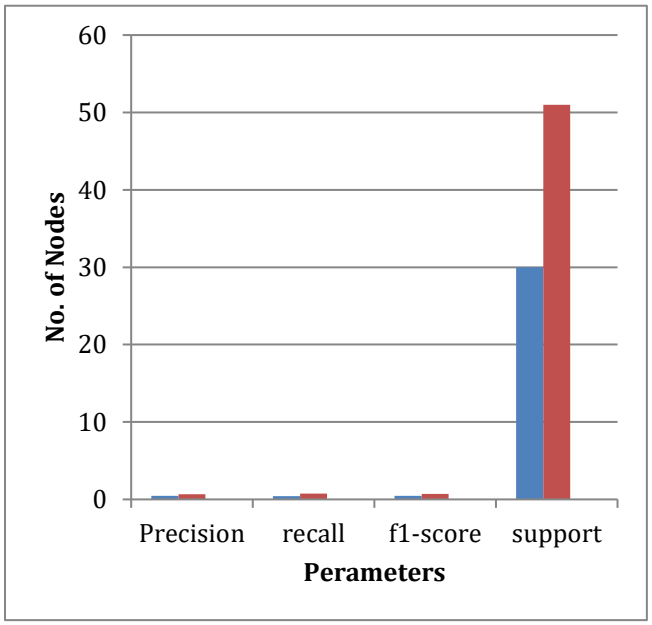

Figure 3: No. of Nodes with Parameters Graph

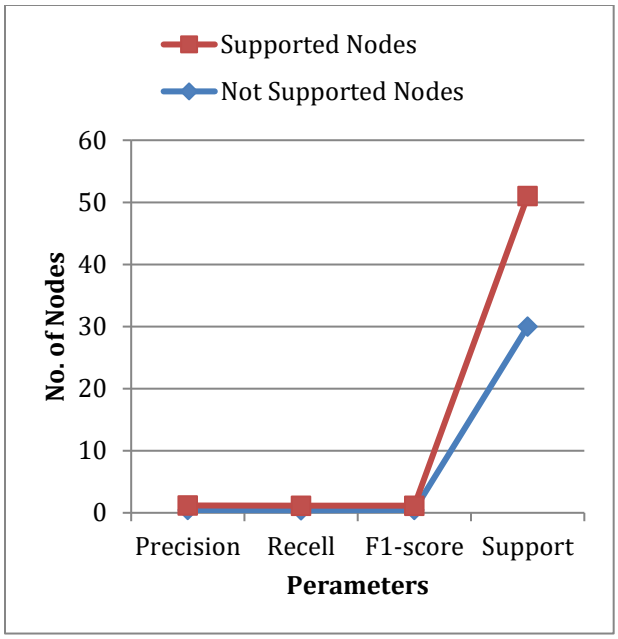

Figure 4: No. of Nodes with supported and not supported graph

Table 2: Confusion Matrix Values with Derivation

Measure	Value	Derivations
Sensitivity	0.7308	TPR = TP / (TP + FN)
Specificity	0.4333	SPC = TN / (FP + TN)
Precision	0.6909	PPV = TP / (TP + FP)
Negative Predictive Value	0.4815	NPV = TN / (TN + FN)
False Positive Rate	0.5667	FPR = FP / (FP + TN)
False Discovery Rate	0.3091	FDR = FP / (FP + TP)
False Negative Rate	0.2692	FNR = FN / (FN + TP)
Accuracy	0.6292	ACC = (TP + TN) / (P + N)
F1 Score	0.7103	F1 = 2TP / (2TP + FP + FN)
Matthews Correlation Coefficient	0.1692	TP*TN - FP*FN / sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))

6. Conclusion

The various limitations of cloud computing have been solved through the use of platform fog computing [1]. By connecting computing, storage, and networking to the network edge, fog helps minimize compute latency, security, network bandwidth, reaction time, connection fees, and power consumption when network resources are limited. To do this, effective fog node services must be discovered and selected as needed. This is a precondition for reaching these goals. The best fog node service should be placed first, followed by the best fog node. To achieve these objectives, certain selection criteria must be satisfied. The goal of this research is to locate the optimal fog node for the service discovery framework that can be used in fog computing. In this study, we proposed a model for a fog computing service discovery framework based on random forest and GridSearchCV. In this study, we provided a model for a fog computing service discovery framework based on GridSearchCV and Random Forest. We apply our suggested machine learning model to the given dataset in this study, analyze the outcomes, and explain the findings. We employed gridsearchcv, logistic regression, random forest, and support vector machine techniques in our

model. To compare the findings, we run a variety of other algorithms on the dataset. Although other algorithms provide more accuracy, we are focusing on random forest and GridSearchCV since they were used with specific purposes and targets. We changed our proposed model using Python environments. Finally, GridSearchCV has the best accuracy of 72%, while Logistic Regression, Support Vector Machine, and Random Forest have the lowest accuracy of 59%, 60%, and 62%, respectively. GridSearchCV has greater in-sample average performance for this sample data, according to this result.

References

[1] Bukhari, Afnan, Farookh Khadeer Hussain, and Omar K. Hussain. "Fog node discovery and selection: A Systematic literature review." *Future Generation Computer Systems* 135 (2022): 114-128.S.

[2] Tang, Tsung-Yi, Li-Yuan Hou, and Tyng-Yeu Liang. "An IOTA-Based Service Discovery Framework for Fog Computing." *Electronics* 10, no. 7 (2021): 844.

[3] Viji Rajendran, V., and S. Swamynathan. "SD-CSR: semantic-based distributed cloud service registry in unstructured P2P networks for augmenting cloud service discovery." *Journal of Network and Systems Management* 27 (2019): 625-646.

[4] Liang, Haoran, Jun Wu, Xi Zheng, Mengshi Zhang, Jianhua Li, and Alireza Jolfaei. "Fog-based secure service discovery for internet of multimedia things: a cross-blockchain approach." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, no. 3s (2020): 1-23.

[5] Sohofi, Ali, Aliasghar Amidian, and Yaghoub Farjami. "SocioChain: A Distributed Ledger-based Social Framework for the Internet of Things." *International Journal of Intelligent Engineering & Systems* 14, no. 5 (2021).

[6] Patil, V. C., K. A. Al-Gaadi, D. P. Biradar, and M. Rangaswamy. "Internet of things (Iot) and cloud computing for agriculture: An overview." *Proceedings of agro-informatics and precision agriculture (AIPA 2012), India* 292 (2012): 296.

[7] Gupta, Brij B., and Megha Quamara. "An overview of Internet of Things (IoT): Architectural aspects, challenges, and protocols." *Concurrency and Computation: Practice and Experience* 32, no. 21 (2020): e4946.

[8] Al-Qaseemi, Sarah A., Hajer A. Almulhim, Maria F. Almulhim, and Saqib Rasool Chaudhry. "IoT architecture challenges and issues: Lack of standardization." In *2016 Future technologies conference (FTC)*, pp. 731-738. IEEE, 2016.

[9] Mohammed, Chnar Mustafa, and Subhi RM Zeebaree. "Sufficient comparison among cloud computing services: IaaS, PaaS, and SaaS: A review." *International Journal of Science and Business* 5, no. 2 (2021): 17-30.

[10] Stojmenovic, Ivan, Sheng Wen, Xinyi Huang, and Hao Luan. "An overview of fog computing and its security issues." *Concurrency and Computation: Practice and Experience* 28, no. 10 (2016): 2991-3005.

[11] Al-Sayed, Mustafa M., Hesham A. Hassan, and Fatma A. Omara. "An intelligent cloud service discovery framework." *Future Generation Computer Systems* 106 (2020): 438-466.

[12] Park, Kisung, Youngho Park, Ashok Kumar Das, Sungjin Yu, Joonyoung Lee, and Yohan Park. "A dynamic privacy-preserving key management protocol for V2G in social internet of things." *IEEE Access* 7 (2019): 76812-76832.

[13] Kumar, Manoj, Dr Mohammad Husain, Naveen Upreti, and Deepti Gupta. "Genetic algorithm: Review and application." *Available at SSRN 3529843* (2010).

[14] Mathew, Tom V. "Genetic algorithm." *Report submitted at IIT Bombay* (2012): 53.

[15] Sarhan, Mohanad, Siamak Layeghy, Nour Moustafa, Marcus Gallagher, and Marius Portmann. "Feature extraction for machine learning-based intrusion detection in IoT networks." *Digital Communications and Networks* (2022).

[16] Park, Kisung, Youngho Park, Ashok Kumar Das, Sungjin Yu, Joonyoung Lee, and Yohan Park. "A dynamic privacy-preserving key management protocol for V2G in social internet of things." *IEEE Access* 7 (2019): 76812-76832.

[17] Sohofi, Ali, Aliasghar Amidian, and Yaghoub Farjami. "SocioChain: A Distributed Ledger-based Social Framework for the Internet of Things." *International Journal of Intelligent Engineering & Systems* 14, no. 5 (2021).

[18] Al-Qaseemi, Sarah A., Hajer A. Almulhim, Maria F. Almulhim, and Saqib Rasool Chaudhry. "IoT architecture challenges and issues: Lack of standardization." In *2016 Future technologies conference (FTC)*, pp. 731-738. IEEE, 2016.

[19] Jabraeil Jamali, Mohammad Ali, Bahareh Bahrami, Arash Heidari, Parisa Allahverdizadeh, Farhad Norouzi, Mohammad Ali Jabraeil Jamali, Bahareh Bahrami, Arash Heidari, Parisa Allahverdizadeh, and Farhad Norouzi. "IoT architecture." *Towards the Internet of Things: Architectures, Security, and Applications* (2020): 9-31.

[20] Al-Sayed, Mustafa M., Hesham A. Hassan, and Fatma A. Omara. "An intelligent cloud service discovery framework." *Future Generation Computer Systems* 106 (2020): 438-466.

[21] Ananthi, M., R. Sabitha, S. Karthik, and J. Shanthini. "FSS-SDD: fuzzy-based semantic search for secure data discovery from outsourced cloud data." *Soft Computing* 24 (2020): 12633-12642.

[22] Liang, Haoran, Jun Wu, Xi Zheng, Mengshi Zhang, Jianhua Li, and Alireza Jolfaei. "Fog-based secure service discovery for internet of multimedia things: a cross-blockchain approach." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, no. 3s (2020): 1-23.

[23] Sharma, Geeta, and Sheetal Kalra. "A lightweight user authentication scheme for cloud-IoT based healthcare services." *Iranian Journal of Science and Technology, Transactions of Electrical Engineering* 43 (2019): 619-636.

[24] Cai, Zhiming, Ivan Lee, Shu-Chuan Chu, and Xuehong Huang. "SimSim: a service discovery method preserving content similarity and spatial similarity in P2P mobile cloud." *Journal of Grid Computing* 17 (2019): 79-95.

[25] Xu, Chenhan, Kun Wang, Peng Li, Song Guo, Jiangtao Luo, Baoliu Ye, and Minyi Guo. "Making big data open in edges: A resource-efficient blockchain-based approach." *IEEE Transactions on Parallel and Distributed Systems* 30, no. 4 (2018): 870-882.

[26] Navimipour, Nima Jafari. "A formal approach for the specification and verification of a trustworthy human resource discovery mechanism in the expert cloud." *Expert Systems with Applications* 42, no. 15-16 (2015): 6112-6131.

[27] Parhi, Manoranjan, Sanjaya Kumar Jena, Niranjan Panda, and Binod Kumar Pattanayak. "A semantic matchmaking technique for cloud service discovery and selection using ontology based on service-oriented architecture." In *Intelligent and Cloud Computing: Proceedings of ICICC 2019, Volume 1*, pp. 31-40. Springer Singapore, 2021.

[28] Abbas, Ghulam, Amjad Mehmood, Jaime Lloret, Muhammad Summair Raza, and Muhammad Ibrahim. "FIPA-based reference architecture for efficient discovery and selection of appropriate cloud service using cloud ontology." *International Journal of Communication Systems* 33, no. 14 (2020): e4504.

Machines and Algorithms
http://www.knovell.org/mna

Research Article

Computer Aided Deep Image Captioning for Medical Images

Hareem Ayesha[1, *], Mehreen Tariq[2, *] and Sundas Israr[3]

[1]Institute of Computer Science & Information Technology, The Women University, Multan, 60000, Pakistan
[2]Department of Computer Science, Bahauddin Zakariya University, Multan, 60000, Pakistan
[3]Department of Computer Science, NUML, Multan, 60000, Pakistan
[*]Corresponding Author: Hareem Ayesha. Email: hareem.ayesha@wum.edu.pk

Received: 18 November 2022; Revised: 22 December 2022; Accepted: 08 February 2023; Published: 6 March 2023
AID: 002-01-000018

Abstract: Drug development, illness diagnosis, prognosis, and prediction are just a few of the many different uses for medical imagery. Ultrasound, X-rays, CT scans, positron emission tomography (PET), and magnetic resonance imaging (MRI) are some of the most common forms of medical imaging. After expert medical professionals have examined these medical photos, they meticulously document their findings in detailed written reports, identifying whether the images are normal, abnormal, or potentially abnormal. This reporting procedure requires a lot of human labor, is prone to mistakes, and takes a lot of time. There have been several proposals for computer-aided report production systems to improve the efficiency and standardization of medical picture reporting. These programs mimic human doctors' work by automatically extracting information from medical images using image captioning techniques and then generating comprehensive written reports. Here, image captioning—which lies at the crossroads of AI's computer vision and natural language processing domains—is crucial. One potential way that medical practitioners may speed up their diagnostic processes is by automating the creation of medical reports. Creating medical reports for chest X-rays is the primary goal of this study's deep learning-based algorithm. A few examples of the many uses for the generated reports include verifying assumptions, keeping tabs on minor adjustments, getting a second opinion, helping with final decisions, getting quick and primary data on the issue under consideration, and much more besides. Therefore, the reports that are automatically created provide critical data for future medical treatments, instead of waiting for a report from an expert doctor.

Keywords: Medical imagery; Computer-aided systems; Image captioning; Deep learning; Diagnostic automation;

1. Introduction

There are numerous applications for medical images in the medical field. From helping chemists discover new drugs to providing surgeons with crucial information during pre-, post-, and intra-operative phases of procedures, these images are everywhere. Radiologists and other medical professionals depend on medical imaging for illness diagnosis and treatment. In the medical field, X-rays, CT-Scans, PET scans, ultrasonography, and MRIs are some of the most common imaging modalities used. After carefully analyzing these medical images, skilled doctors write extensive reports that summaries their findings. These reports are given in paragraph form and can be categorized as normal, abnormal, or potentially abnormal.

https://doi.org/10.66108/mna.v2i1.36

Writing medical reports from scratch requires an in-depth familiarity with the condition, medical imaging, and thorough review of the pictures; this can be especially difficult for novice examiners. It takes at least thirty minutes to review each picture and write up the results, which is a lot of time even for experienced doctors. This task is made worse in areas where medical practitioners are in short supply, like Pakistan, where the population is large, which increases the chances of wrong diagnosis [1]. In low-income nations like Pakistan, the situation is worsened since it costs more to follow up with physicians for report-related inquiries.

To make medical image reporting easier, many computer-aided report-generating systems have been suggested, all based on picture captioning. Like a seasoned doctor, these computers can automatically analyze medical photos for results and write out detailed reports. Medical professionals can save time and avoid employing extra staff to write reports manually thanks to this technology. Radiologists can use the reports for monitoring and cross-checking, getting a second opinion from other doctors, and giving techs quick information, among other uses. These reports are automatically created and provide crucial context for starting treatment quickly in situations of emergency when experienced doctors may not be easily accessible.

Although there are clear benefits, medical picture captioning does present a number of obstacles. The process of writing an extensive paragraph for a medical report is far from easy compared to making a caption consisting of only one phrase. In addition to various kind of textual descriptions, medical reports also include impressions (diagnoses), comparisons, and keyword tags generated from important discoveries. A number of steps are required to tackle these intricacies, including segmentation, feature extraction, classification, pre-processing (to reduce noise in medical pictures), and visual feature selection. Finding abnormality zones using segmentation is difficult in general, but especially when dealing with noisy modalities like ultrasound [2]. Furthermore, overfitting and possible discrepancies between produced and original captions or reports are consequences of the lack of high-quality datasets for medical picture captioning, which in turn hinders model generalization. Making sure the produced medical reports are accurate, legible, and free of grammar and spelling errors just adds to the difficulties already encountered in this field.

1.1. Objectives of Research

The primary goal of this thesis is to create a deep learning (DL) model for generating radiology reports from medical images. Other objective of this research is described below:

- Investigate the diverse applications of DL in the medical field, including image classification, segmentation, translation, retrieval, and disease detection. With particular attention to how important it is for DL to automate disease prognosis and diagnosis across all modalities so that medical professionals may make better decisions.
- Conduct a comprehensive review of existing literature to identify gaps and challenges in current methods. Examine the most important factors affecting how well automatic medical report generating works.
- Develop a DL-based model specifically designed for generating grammatically correct medical reports based on medical images.

2. LITERATURE REVIEW

The utilization of deep learning networks to accomplish natural language-like image captioning in recent years [25,26] has generated enthusiasm for the application of deep learning methodologies in the domain of medical image captioning. Existing literature frequently employs the encoder-decoder architecture, in which a CNN encodes and decodes image characteristics into fixed-length vector representations. Subsequently, the encoder Recurrent Neural Network (RNN) such as LSTM or Gated Recurrent Units (GRU) is supplied with these representations; it generates word sequences that possess precise syntax and elaborate semantics. The training process for these models commences and concludes, eliminating the

necessity to align individual modules. One drawback of this technique is that it combines multiple categories of retrieved image characteristics into a single fixed-length vector representation, which is then utilized by the decoder.

This endeavor centers around the creation of captions and textual labelling. The subsequent section presents an exhaustive analysis of caption generating methods based on deep learning. Figure 1 illustrates the taxonomy of medical image captioning systems that utilize DL.

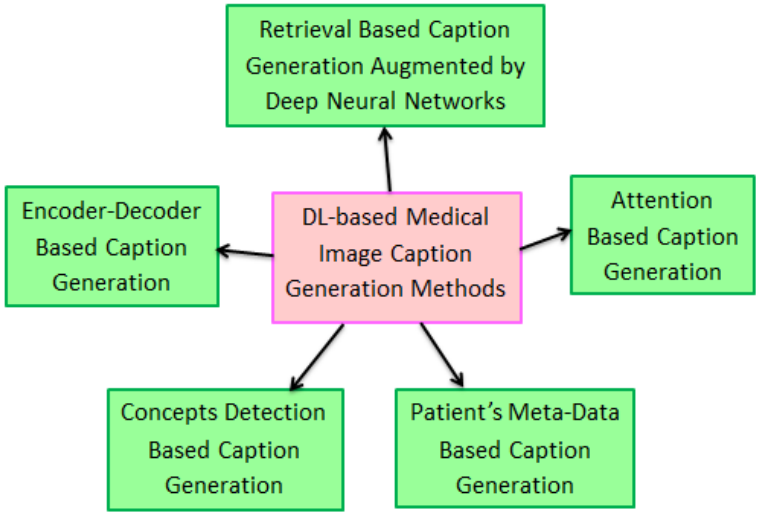

Figure 1: DL based medical image caption generation methods

2.1. Encoder-decoder based caption generation

In order to create their NN-based algorithms, the examined study makes use of Transfer Learning (TL). An encoder-decoder architecture works by first using a model that has been pre-trained on real-world pictures and then refining it using data from a specific domain. With the exception of Lyndon et al. [19], all of the other studies employ single-layer LSTM decoders. To prevent overfitting, regularization was utilised by Shin et al. [5], Wu et al. [10], Su et al. [16], and Lyndon et al. [19]. No regularisation approach was mentioned in the studies of Zeng et al. [2], Pelka et al. [18], and Spinks et al. [24]. All of these strategies provide brief captions. Despite its success, the captions produced by Shin et al. [5] are not a cohesive report but rather a 5-word "bag of words" that just explain the environment of a defined condition. The captions produced in the study by Wu et al. [10] solely included picture anomalies. Since the only focus of Wu et al. [10] was on the produced caption's relationship to the picture, specificity, sensitivity, and accuracy are the evaluation metrics utilised. Even though they included regularisation in their model, the dataset was too tiny to rule out the possibility of overfitting. In their study, Liang et al. [12] demonstrated that training on multiple models using different sub-datasets with different properties and finally classifying through SVM improved performance. However, their model is not well-suited for producing complex and fully descriptive captions, such as natural language descriptions. In the absence of a complete sentence caption, Pelka et al. [18] likewise produced just keywords. Faster RCNN was employed by Zeng et al. [2] to accomplish both picture area identification and encoding inside a single model. It outperformed captioned methods used for ultrasound pictures, which take full-size images into account, and two independent models for detection and encoding. Their model was more efficient and required fewer parameters, saving time. Nevertheless, this approach does have a few drawbacks, such as the fact that it generates brief explanations and makes mistakes in language and word class prediction when creating captions.

Table 1: Encoder Decoder based caption generation.

Authors	Encoder	Additional Processing	Decoder	Transfer Learning	Regularization
[5]	GoogleNet	Image classification and detection	LSTM, GRU	Yes (fine-tuning)	Dropout, Batch normalization
[2]	VGG-16	Region detection, Classification Regression	LSTM	Yes (fine-tuning)	N/A
[10]	Batch Normalized CNN [17]	No	LSTM	Yes	Dataset Expansion, Dropout
[19]	Inception-V3	Concepts detection	3- layer LSTM	Yes	Dropout
[12]	VGGNet	SVM classification	LSTM	Yes	N/A
[18]	Inception-V3	No	LSTM	Yes (fine-tuning)	N/A
[16]	ResNet-152	Experiment on VGGNet	LSTM	Yes (fine-tuning)	Dataset Expansion
[24]	N/A	Coding Images into continuous representation	N/A	N/A	Early stopping

2.2 Attention based caption generation

Our evaluation revealed that the majority of research models for captioning are constructed utilising transfer learning. As a rule, these models make use of pre-trained encoders that have been either fine-tuned on domain datasets or utilised without any fine-tuning at all; however, in the case of Zhang et al. [4], the authors train their own encoder from the ground up. A comprehensive radiology diagnostic report is more difficult to develop than the five sorts of bladder characteristics described in the reports produced by Zhang et al. [4] based on cellular appearance (Fig. 6). Jing et al. [3] is the only paper in the attention-based caption creation category that calculates semantic attention over extracted information; all other papers in this category compute attention over solely visual spatial features. The majority of the research presented here found that employing sentence-LSTM and word-LSTM together as a decoder enhanced performance. Jing et al. [3] used hierarchical LSTM to get decent results, although the reports they made use of repeated terms. It is possible that these repeats are caused by their hierarchical model ignoring contextual coherence. Similar to Jing et al. [3], Wang et al. [9] conducted experiments on the OpenI dataset; however, they did not provide the final assessment findings. By examining the produced reports and concentrating on their experiment with the Chest X-Rays dataset, it is evident that the findings are inferior to the OpenI results reported by Jing et al. [3]. The usage of flat LSTM to decode the textual reports could be the cause of this. The reports produced by Xue et al. [15] are well-organized, although they may benefit from fewer repeats. Additionally, some of the produced reports do not catch all of the irregularities. Training the model on a limited dataset with a small number of aberrant samples might be the reason of this erroneous behaviour. It is not easy to develop well-formed sentences when they lack underlying facts. The difficulty of acquiring grammatical accuracy from tiny samples is another possible explanation. The authors Gale et al. [6] were unable to pinpoint the exact site of the fracture, but their model was able to convey the fracture's characteristics in straightforward language. According to Yuan et al. [17], combining fusion mechanism and concept information with decoding process enhances performance, with late fusion outperforming early

fusion. Training an encoder using a dataset that is particular to a domain improves its performance. Unfortunately, their model isn't superior at producing unseen words because of the tiny scale dataset.

Table 2: Attention based caption generation.

Authors	Encoder	Additional Processing	Type of Attention	Decoder	Transfer Learning	Regularization
[3]	VGG-19	Multi-label Classification to predict tags	Visual spatial and semantic	Hierarchical LSTM (Sentence ,Word)	Yes (fine-tuning)	Yes (Early Stopping)
[4]	Own designed new ResNet	Symptom description-based image retrieval	Visual spatial	LSTM	No	Yes (Gradient Optimization)
[15]	ResNet-152	--	Visual attention over image and semantic attention over sentence	Hierarchical BiLSTM	Yes (Pre-trained)	--
[6]	DenseNet	Manual labeling of images	Visual spatial	BiLSTM	Yes (Pre-trained)	Yes (Dropout , augmentation)
[9]	ResNet-50	Auto-annotation and classification of images	Visual spatial	LSTM	Yes (Pre-trained, fine-tuning)	Yes (Dropout, L2 regularization)
[17]	ResNet-152	Classification of chest radiographic observations	Visual spatial	Hierarchical LSTM (Sentence ,Word)	Yes (pre-trained on CheXpert dataset)	--
[14]	--	Disease classification	Visual spatial	Hierarchical LSTM (Sentence ,Word)	--	--
[13]	VGGnet-19	Concepts generation	Visual spatial	LSTM	Yes (fine-tuning)	Yes (Dropout)
[20]	ResNet-101	Concepts detection	Visual spatial	LSTM	Yes (Pre-trained)	Yes (Dropout, early stopping)
[21]	RseNet-151	Disease classification, localization	Visual spatial	LSTM	Yes (Pre-trained)	--

2.3. Patient's meta-data Based Caption Generation:

A medical report's background or indication part will often go over patient information, symptoms of the ailment, and past therapies. You may include this data by merging a training model with the results section, but you'll need to train your model well to tell them apart. In order to circumvent this overhead and

direct the decoder to provide more precise results, the patient's background information is encoded independently.

Incorporating background information for report generation requires little effort and yields good results. While Zhang et al. [23] did employ regularisation and transfer learning in their model, they only produced an impression sentence and did not take medical images as input. Instead, they used a results paragraph. Rather than focusing just on visual attention, Huang et al. [11] calculated spatial and channel attention. The resulting report differed greatly from the original report, though, since the dataset was too tiny. Details of research that used patient meta-data in their reviews The results may be seen in Table 5.

Table 3: Patient's meta-data Based Caption Generation.

Authors	Encoder	Additional Processing	Decoder	Transfer Learning	Regularization
[11]	ResNet-152	Spatial and Channel Attention	Hierarchical BiLSTM (Word ,Background, Sentence LSTM)	Yes	Yes (Dropout)
[23]	BiLSTM	Attention mechanism	LSTM	No	--

2.4. Retrieval based caption generation augmented by deep neural networks

Generate new descriptions for input photos using pre-existing captions in a database using retrieval-based image caption creation. Captions that aren't related to the scene or item are the result of this method's inability to adapt. Scientists are working to enhance its capabilities by merging retrieval-based caption creation with deep neural networks.

While a combination of retrieval-based captioning and deep neural networks can improve performance, there is still room for improvement when it comes to medical report generation and medical picture captioned. Liang et al. [12] found that while results may vary depending on the dataset used to train the model, utilising a combination of support vector machines (SVMs) and a convolutional neural network (CNN) improves performance, and making use of the extracted caption can yield even better results. However, complicated captions that are entirely descriptive will not be generated using the suggested paradigm. Li et al. [22] suggested a strategy where RNN-derived captions linked to sentence topics outperform new captions created using encoder-decoder architecture. However, they also noted that obtaining captions from template corpora does not adequately describe some unusual discoveries. However, with great accuracy, the caption creation module produces captions that include aberrant results. In their papers, none of them mentioned regularisation.

Table 4: Summarises the details of the research that used this approach for caption generating.

Authors	Encoder	Additional Processing	Decoder	Transfer Learning	Regularization
[7]	GoogleNet for multi-label classification	Concepts prediction	--	Yes (fine-tuning)	--
[12]	VGGNet	SVM classification	LSTM	Yes	--
[22]	DenseNet	Reinforcement learning, attention mechanism	Hierarchical RNN	Yes (fine-tuning, pre-trained)	--
[8]	Inception-V3	LIRE retrieval System	--	Yes (pre-trained)	--

2.5. Concepts detection Based Caption Generation

Using the visual look of medical photographs as a starting point, this technique aims to develop medical concepts that may then be used as captions. These ideas are seen as separate components that may be used to create captions.

Little work has been discovered on producing medical captions using this approach, similar to other categories. While several studies made advantage of transfer learning, only Hasan et al. [13] developed a model that included an attention mechanism. Their better score compared to Ben abacha et al. [7] might be explained by this. The ICLEFcaption2017 competition includes both of these works. Despite using the same dataset, they omitted information about the dataset's pre-processing from their publications. In their regularisation approach, Hasan et al. [13] utilised dropout, however Ben abacha et al. [7] made no mention of regularisation at all. Unfortunately, neither author has produced a comprehensive medical report, and the captions they have produced are just one or two sentences long. Table 7 provides a summary of the research that were examined and used concepts detection-based caption creation.

Table 5: Using deep neural networks to enhance caption production based on concepts detection

Authors	Encoder	Additional Processing	Decoder	Transfer Learning	Regularization
[7]	GoogleNet	Concepts detection	--	Yes (fine-tuning)	--
[13]	VGGnet-19	Visual soft attention	LSTM	Yes (fine-tuning)	Yes (Dropout)

3. Research Methodology

The subsequent approach was implemented in order to optimize the process of medical image captioning.

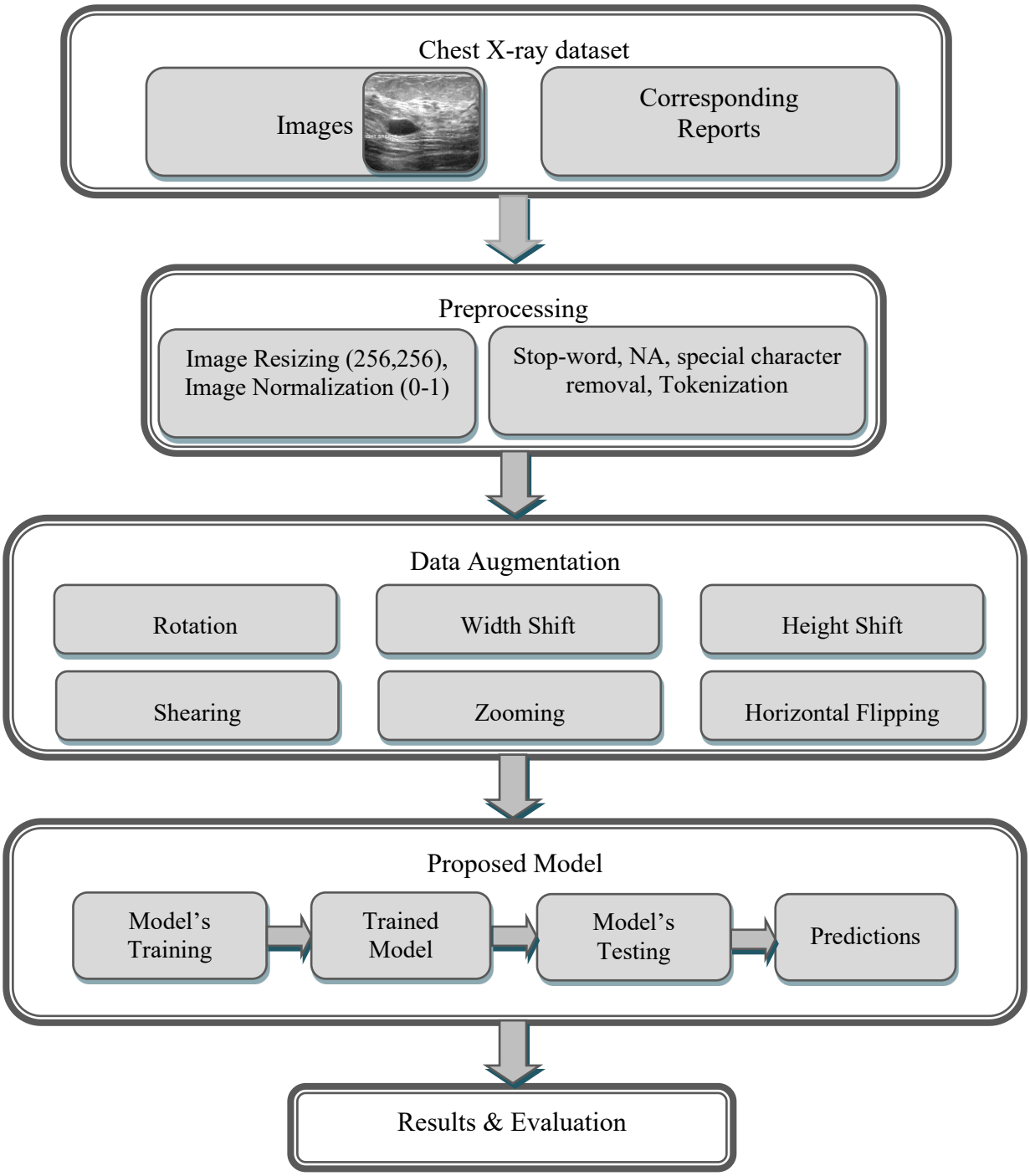

Figure 2: Research Methodology

3.1. Dataset Description

Through OpenI, we were able to obtain the IU Chest X-ray dataset, which contains 7,470 pictures of the chest and its sides, together with 3,955 radiology reports. Each image is linked to a radiology report comprising five sections: Impression (final diagnosis), Comparison (previous treatment details), Indication

(patient symptoms and meta-data), and Findings (radiologists' observations). The Tags section contains keywords derived from critical information in Impression and Findings, manually encoded with MeSH terms and auto-encoded using MTI. These tags are vital for generating terms in the final caption process. Despite researchers often using only Impression and Findings, our generated descriptions are more detailed [3].

3.2. Dataset Preparation and Preprocessing

Data preparation is a key step in developing an efficient, high-quality, and competitive model. Each algorithm has specific prerequisites that must be met in order to produce the intended results. As a result, data is preprocessed before being assigned to an algorithm. This study's data preparation and preprocessing is detailed below.

3.2.1. Image Preprocessing

In the preprocessing phase, addressing variations in image sizes within the selected dataset was essential for compatibility with deep learning networks. Two strategies were considered: padding, involving the addition of extra columns and rows to achieve uniform size but potentially incurring additional computational costs, and image resizing, where images were rescaled to a standardized (256*256) pixel size, reducing computational overhead. Image normalization was used to achieve optimal performance throughout model training. Deep model weights are often initialized with tiny random values (0-1), but the intensity range of input pictures is greater (0-255). It was essential to resize input images to a normalized range of 0–1 in order to minimize problems like bursting gradients that can impair learning.

3.2.2. Report Preprocessing

In the process of preparing the dataset for caption generation, XML reports downloaded from OpenI were initially converted into an Excel format, with only essential details extracted for input to the model. Subsequently, certain preprocessing steps were applied to refine the textual data. 'xxx' or 'xxxx', used to replace patient information in the original X-ray reports before public release, were deemed irrelevant for our purposes and thus removed. Additionally, all punctuation marks, special characters, and digits (0-9) were eliminated. Reports lacking a findings section were excluded from the dataset. To ensure consistency, all reports were converted to lowercase. Special tokens, 'startseq' and 'endseq', were introduced at the beginning and end of captions, aiding the model in recognizing sentence boundaries. Tokenization was employed, breaking down the entire report into words, with each word assigned a unique numerical token. Every unique word was kept in an array called vocabulary, and each word had a token number assigned to it. Since deep learning algorithms need numerical data, word embedding was done outside the model by giving each word a distinct token number and generating a word embedding matrix that could be integrated into the model.

3.3. Data Augmentation

During the training of deep learning architectures, a prevalent challenge is overfitting, where the model memorizes input data, yielding better results during training compared to testing. This problem is frequently caused by a lack of training data, which is a prevalent scenario in the automated diagnosis of numerous diseases due to the scarcity of large-scale medical imaging datasets. In order to reduce overfitting, we present a methodology that uses data augmentation. Specifically, we perform six different modifications to enlarge the dataset:

- Rotation: Images are rotated at random within a 0.2-degree range.
- Width Shift: Input photos are randomly shifted horizontally to the left or right by 0.05% of their entire width.
- Height Shift: Random vertical changes uphill or downward within a range of 0.05% of the overall height occur in input photos.
- Shearing: Input pictures are sheared anticlockwise within a 0.05-degree range.

- Zoom: The input photos are zoomed at random within the range [1-0.05, 1+0.05].
- Horizontal Flipping: Input pictures are flipped horizontally at random.

3.4. Proposed Architecture

The proposed architecture is divided into two distinct categories. A CNN performs multi-label classification of chest x-rays as the initial model. LSTM is utilized as a captioning algorithm in the second section to generate textual paragraphs. The diagram below illustrates the strategic architecture.

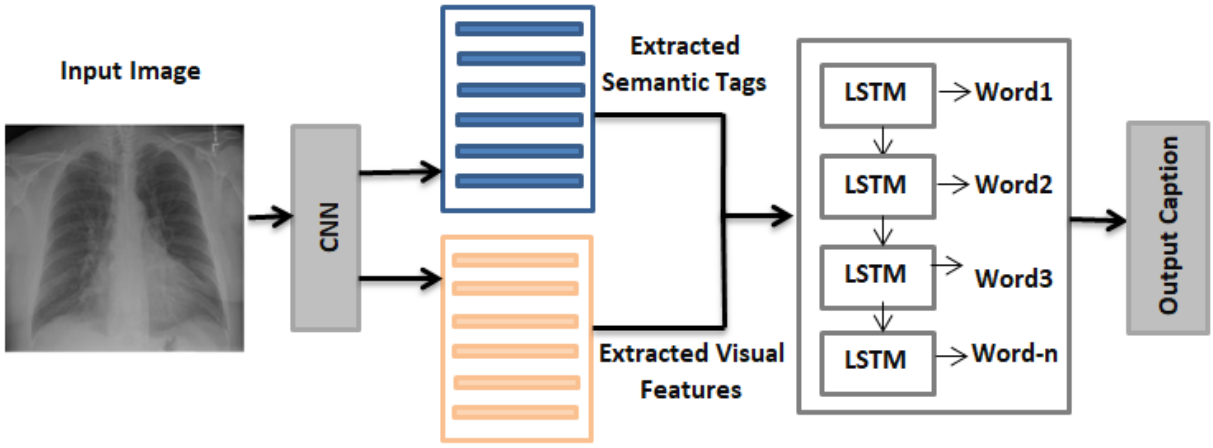

Figure 3: Proposed architecture

3.4.1. Multi Label Classification (MLC)

Convolutional neural networks (CNNs) clearly shine in binary or multi-label classification situations, according to the current research. The Keras programme allows several CNN variants to save their weights for future use. These variations are trained using large picture datasets. Researchers love these pre-trained models because they are easy to use, improve performance, and cut down on computing time. We used ChexNet [6], a pre-trained network that is indicative of its kind, in our investigation.

Specifically trained on the 14-class ChestXray14 dataset, ChexNet is a convolutional neural network. To detect 14 anomalies from 112,120 ChestX-rays, ChexNet was built using a 121-layer architecture. To make tag predictions, we used this model for both feature extraction and multi-label classification. In this case, given a picture, the task was to extract characteristics from the second-to-last layer. To make the model work for tag prediction, we added 210 nodes to the last dense layer and treated it as an MLC object. We used the top ten tags, which were determined to be the most semantically meaningful by the MLC model, to train our caption generating model. The model was trained from the ground up using a dataset of 6512 lateral and frontal chest X-rays for feature extraction and tag prediction.

3.4.2. Caption Generation

With the use of ChexNet features, an LSTM language model can create captions. Following feature extraction, a 256-unit LSTM is supplied with a word embedding matrix. In order to generate new output, the LSTM looks at previous ones. The caption model takes three things into consideration while creating a word-by-word caption: picture characteristics, extracted tags, and LSTM output.

3.5. Training Parameters

Image types in the original dataset were 70% training, 20% validation, and 10% testing. ChexNet trained with batch size 10, 100 epochs, Adam optimizer (lr=1e-4, binary_crossentropy). Optimal model weights preserved according to validation loss. CNN weights that have been trained are used for tag prediction and

feature extraction. Language model trained on 70% dataset with batch size 10, 100 epochs, Adam optimizer (lr=1e-4, categorical_crossentropy). Both models in the proposed architecture underwent image augmentation.

4. Results and Evaluation

In this chapter, we assess the outcomes of the deep learning-based architecture that we have proposed for the automated generation of captioning pertaining to chest X-rays. The experimental procedures and the system utilised in this study are detailed below:

4.1. Experimental Analysis

Two models for Multi-Label Classification (MLC) and tag prediction were trained, using several training and testing rounds to optimise parameters for our unique scenario. The model was trained in the first experiment by replacing the final layer of the pre-trained ChexNet with 210 nodes, which corresponded to the 210 tags in our research. ChexNet was trained from scratch on our chosen dataset for 80 epochs in the second experiment, with continuous monitoring of training and validation loss. Weights were saved for both models after effective training. These weights were loaded during testing, and tags were expected. The model created from scratch beat the pre-trained model, which served as the encoder in our design, according to the results. The decoder gathered features and tags from this model in order to create the final caption. Using the training datasets, the whole network was trained across 80 epochs.

4.2. Training Loss and Validation Loss

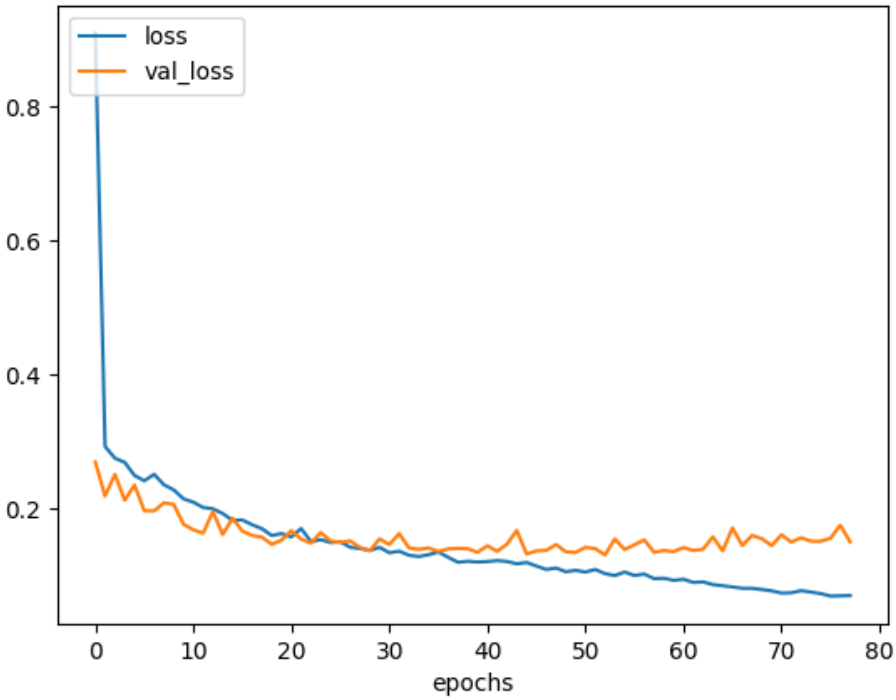

Figure 4: Training Loss and Validation Loss of Pre-trained ChexNet

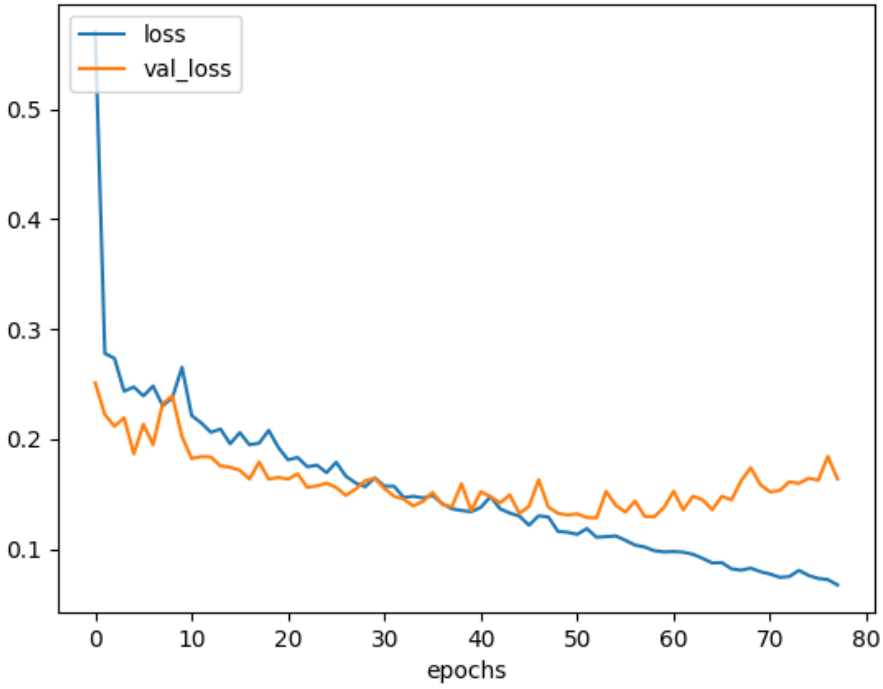

Figure 5: Training Loss and Validation Loss of ChexNet Trained from scratch

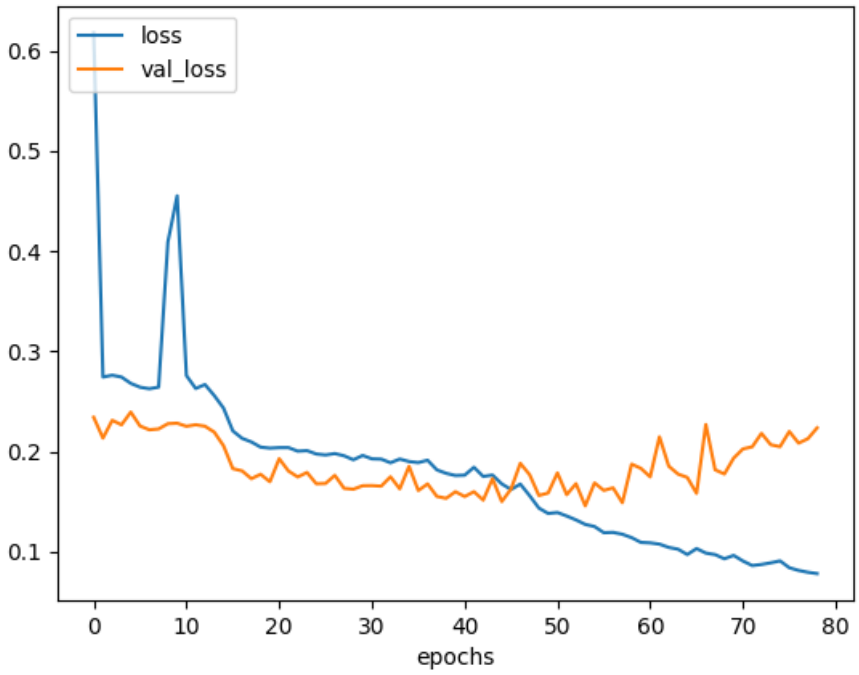

Figure 6: Training Loss and Validation Loss of Whole Network without Tags

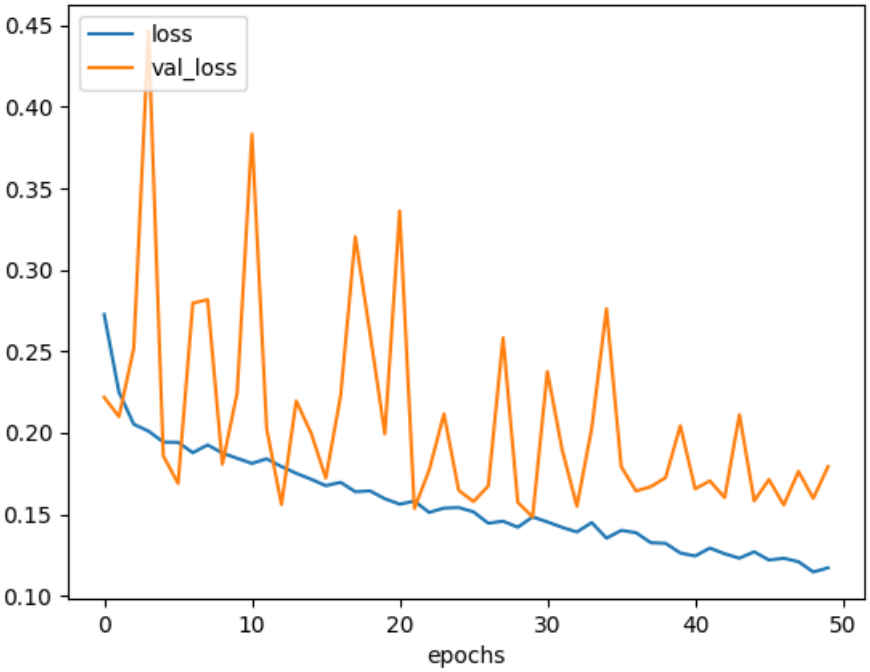

Figure 7: The Impact on Training and Validation Deletion of Entire Network Using Tags

4.2.1. Predictions of Proposed Model

Below are the figures displaying the projected reports and their matching original reports that were produced from our suggested model using 2 test images:

Original Report	Predicted Report
the lungs appear clear the heart and pulmonary are normal the pleural spaces are clear mediastinal contours are normal	lungs appear clear the heart is normal the pleural spaces are clear mediastinal contours are normal

images heart size and pulmonary vascular engorgement appear within limits of normal mediastinal contour is unremarkable no focal consolidation pleural effusion or pneumothorax identified no convincing acute bony findings	Heart size and pulmonary vascular engorgement within limits of normal mediastinal contour is unremarkable no focal consolidation pleural effusion or pneumothorax identified no convincing acute bony findings findings findings or pneumothorax

Figure 8: Predictions from the Proposed Model on two test Images

4.2.2. BLEU Scores

Table 6: Experimental BLEU score and the final model proposal

Dataset	Method	BLEU1	BLEU2	BLEU3	BLEU4
IU Chest X-ray	Pre-trained Encoder	0.212	0.110	0.034	0.033
	From scratch trained Encoder	0.256	0.130	0.056	0.047
	With tags	0.307	0.296	0.327	0.282

The table above displays the BLEU Scores of the four models in relation to the predictions of the Chest X-ray dataset. The "startseq" and "endseq" elements are eliminated from the final prediction produced by each of the four architectures before this score is computed. We outperformed the other two studies and attained a higher BLEU score with the implementation of our proposed design.

CONCLUSION

The integration of artificial intelligence has brought about a transformative impact on the medical field, offering various applications to support medical practitioners in tasks such as automated disease detection and diagnosis, real-time patient monitoring, automated report generation, automated surgeries, and administrative responsibilities. Among these applications, automated report generation or image captioning for different diseases stands out as a particularly challenging endeavor. Unlike simple object recognition, segmentation, or classification, this task necessitates a comprehension of the relationships between different items in a picture and the behaviors that these objects represent. Traditionally, medical picture analysis and the detection of numerous types of anomalies, as well as manual report preparation, take a significant amount of time and effort. Medical image captioning emerges as a solution to address these challenges, aiming to save experts' time, mitigate subjectivity errors, and ensure accurate interpretation of both major and minor abnormalities. The resulting reports can serve as valuable resources for a range of subsequent tasks.

There are two main categories of medical image captioning models. One type uses deep learning to generate captions using different types of neural networks. The other type uses retrieval to find the most similar image-caption pairs and then associates the best one with the input image. Here, we provide a state-of-the-art deep encoder-decoder architecture and use a deep learning-based approach to build captions for medical images. The encoder in the proposed model is composed of two parts: a pre-trained deep features extractor based on the ChexNet network and the extraction of the top ten tags from the Convolutional Neural Network (CNN)'s fully connected layer. These tags will almost certainly act as the encoder's second input. The second section includes a Recurrent Neural Network (RNN), which uses the sequence created up to the (n-1) step as input at the nth step. Following each of the two input layers, a dense layer is added, and the outputs from all three portions are concatenated and given to the decoder. A completely linked hidden layer and an output layer compose the decoder. The proposed model was trained and evaluated using the publicly available chest X-ray dataset from Indiana University. Our methodology received a BLEU1 score of 0.307, confirming its efficacy in the area of medical picture captioning.

Future Work

In the future, we want to use LSTM and an attention mechanism in the proposed network's decoder.

References

[1] Brady, Adrian, Risteárd Ó. Laoide, Peter McCarthy, and Ronan McDermott. "Discrepancy and error in radiology: concepts, causes and consequences." *The Ulster medical journal* 81, no. 1 (2012): 3.

[2] Zeng, Xianhua, Li Wen, Banggui Liu, and Xiaojun Qi. "Deep learning for ultrasound image caption generation based on object detection." *Neurocomputing* 392 (2020): 132-141.

[3] Jing, Baoyu, Pengtao Xie, and Eric Xing. "On the automatic generation of medical imaging reports." *arXiv preprint arXiv:1711.08195* (2017).

[4] Zhang, Zizhao, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. "Mdnet: A semantically and visually interpretable medical image diagnosis network." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6428-6436. 2017.

[5] Liu, Xinglong, Fei Hou, Hong Qin, and Aimin Hao. "Multi-view multi-scale CNNs for lung nodule type classification from CT images." *Pattern Recognition* 77 (2018): 262-275.

[6] Rajpurkar, Pranav, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." *arXiv preprint arXiv:1711.05225* (2017).

[7] Cai, Yiheng, Yuanyuan Li, Changyan Qiu, Jie Ma, and Xurong Gao. "Medical image retrieval based on convolutional neural network and supervised hashing." *IEEE access* 7 (2019): 51877-51885.

[8] Qayyum, Adnan, Syed Muhammad Anwar, Muhammad Awais, and Muhammad Majid. "Medical image retrieval using deep convolutional neural network." *Neurocomputing* 266 (2017): 8-20.

[9] Azam, Sheikh Shams, Manoj Raju, Venkatesh Pagidimarri, and Vamsi Kasivajjala. "Q-Map: clinical concept mining from clinical documents." *arXiv preprint arXiv:1804.11149* (2018).

[10] Soldaini, Luca, and Nazli Goharian. "Quickumls: a fast, unsupervised approach for medical concept extraction." In *MedIR workshop, sigir*, pp. 1-4. 2016.

[11] Liu, Guanxiong, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. "Clinically accurate chest x-ray report generation." In *Machine Learning for Healthcare Conference*, pp. 249-269. PMLR, 2019.

[12] Bustos, Aurelia, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. "Padchest: A large chest x-ray image dataset with multi-label annotated reports." *Medical image analysis* 66 (2020): 101797.

[13] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818-833. Springer International Publishing, 2014.

[14] Zhang, Yu, Xuwen Wang, Zhen Guo, and Jiao Li. "ImageSem at ImageCLEF 2018 caption task: Image retrieval and transfer learning." In *CLEF CEUR Workshop, Avignon, France*. 2018.

[15] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation." In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311-318. 2002.

[16] Banerjee, S., and A. Lavie. "Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments." *Proceedings of ACL-WMT*: 65-72.

[17] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." In *Text summarization branches out*, pp. 74-81. 2004.

[18] Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566-4575. 2015.

[19] Anderson, Peter, Basura Fernando, Mark Johnson, and Stephen Gould. "Spice: Semantic propositional image caption evaluation." In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pp. 382-398. Springer International Publishing, 2016.

[20] Wang, Xiaosong, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9049-9058. 2018.

[21] Wu, Luhui, Cheng Wan, Yiquan Wu, and Jiang Liu. "Generative caption for diabetic retinopathy images." In *2017 International conference on security, pattern analysis, and cybernetics (SPAC)*, pp. 515-519. IEEE, 2017.

[22] Huang, Xin, Fengqi Yan, Wei Xu, and Maozhen Li. "Multi-attention and incorporating background information model for chest x-ray image report generation." *IEEE Access* 7 (2019): 154808-154817.

[23] Hasan, Sadid A., Yuan Ling, Joey Liu, Rithesh Sreenivasan, Shreya Anand, Tilak Raj Arora, Vivek V. Datla et

al. "PRNA at ImageCLEF 2017 Caption Prediction and Concept Detection Tasks." In *CLEF (working notes)*. 2017.

[24] Frangi, Alejandro F., Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-Lopez, and Gabor Fichtinger. "Lecture Notes in Computer Science: proceedings of the Medical Image Computing and Computer Assisted Intervention-MICCAI 2018." (2018).

[25] Kilickaya, Mert, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. "Re-evaluating automatic metrics for image captioning." *arXiv preprint arXiv:1612.07600* (2016).

[26] Elliott, Desmond, and Frank Keller. "Comparing automatic evaluation measures for image description." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 452-457. 2014.

[27] Zeng, Xianhua, Li Wen, Banggui Liu, and Xiaojun Qi. "Deep learning for ultrasound image caption generation based on object detection." *Neurocomputing* 392 (2020): 132-141.

Machines and Algorithms
http://www.knovell.org/mna

Research Article

Computer Aided Diagnosis System for Breast Cancer Detection

Mehreen Tariq[1, *], Hareem Ayesha[2]

[1]Department of Computer Science, Bahauddin Zakariya University, Multan, 60000, Pakistan.
[2]Institute of Computer Science and Information Technology, The Women University, Multan, 60000, Pakistan.
[*]Corresponding Author: Author's Name. Email: mehreentariq97@gmail.com

Received: 23 November 2022; Revised: 23 December 2022; Accepted: 17 February 2023; Published: 6 March 2023
AID: 002-01-000019

Abstract: Recent developments in medical imaging have substantially enhanced the capacity to identify early-stage breast cancer. Medical practitioners frequently employ radiographic and microscopic imaging methods to detect a wide range of breast disorders. Accurately classifying these anomalies is challenging because of the similarities in appearance between benign and malignant breast cancers as well as the imprecision of imaging data. The integration of AI into healthcare applications has stoked the fires of research into intelligent computer-aided detection and diagnostic systems. The combination of computer vision (CV) and image processing allows these systems to identify breast cancer effectively. Despite the abundance of literature on the subject, most of it has focused on treating breast mammography as a binary test for benign or malignant abnormalities. Machine learning-based techniques have restricted accuracy due to their insufficient feature extraction capabilities, despite extensive exploration. However, because of their capacity to extract hundreds of characteristics, deep learning-based methods have demonstrated better performance in several experiments. A prevalent method in these examinations has been transfer learning utilizing pre-trained deep models. In order to semantically detect breast abnormalities in ultrasound pictures, our research presents a new deep architecture. Our suggested model demonstrated outstanding efficiency when compared to three top-tier deep semantic segmentation architectures: UNet, UNet++, and SegNet. We tested our model's generalizability using UDIAT, another publicly accessible BUS imaging dataset, in addition to training it on 546 pictures from a publicly available breast ultrasound imaging dataset. Our model's remarkable performance in semantic breast lesion identification was highlighted by its Jaccard scores of 0.77 on Dataset 1 and 0.748 on Dataset 2.

Keywords: Medical Imaging; Breast Cancer Detection; AI in Healthcare; Deep Learning in Medical Imaging; Computer-Aided Diagnosis;

1. Introduction

Many of areas of medical practice have been impacted by the revolutionary rise of artificial intelligence (AI). Applications in the medical industry have experienced a major spike due to the fast development of new medical data sources and the growth of AI-based technology. Automated computer-aided systems are being developed to automate many processes in healthcare using a variety of artificial intelligence (AI) technologies, such as rule-based expert systems, deep learning (DL), and natural language processing (NLP). Predicting illness risks, detecting and diagnosing diseases, generating reports, monitoring patients,

https://doi.org/10.66108/mna.v2i1.37

operating on patients, providing virtual support, and doing administrative duties are all part of this scope of work.

A big and very successful area among the many AI applications in healthcare is automated illness detection [1]. Not only can the use of AI-based systems in healthcare assist reduce the occurrence of pharmaceutical mistakes such medical picture misinterpretation [2], but it also allows for more accurate and prompt patient treatment, which in turn improves survival rates for potentially deadly diseases. Artificial intelligence also helps doctors by cutting down on unnecessary expenses and saving them time. People living in impoverished regions are even more able to get healthcare as a result.

A plethora of CAD systems have been created to help in the detection of several devastating diseases, including as those affecting the cardiovascular system, the nervous system, and the lungs. These systems also seek to diagnose fungal and viral diseases, such as Ebola and Coronavirus, and to identify malignancies in their early stages, including those of the brain, lungs, breast, prostate, and colorectal areas. The necessity to develop a deep learning-based system that can efficiently use a collection of film-mammography breast pictures for early-stage automated diagnosis of breast cancer is what drives this project.

The World Health Organization (WHO) reports that, worldwide, breast cancer ranks second among all cancers followed by lung cancer, making it a serious public health problem. Stages III or IV are the most common outcomes for breast cancer diagnoses in Pakistan, accounting for more than half of all cases [23]. About 90,000 new cases are recorded in Pakistan each year, with 40,000 people losing their lives as a direct result [21]. Delays in medical consultations, caused by variables including low levels of education, widespread poverty, and general lack of knowledge, are a major contributor to the high death rate [22]. The survival rate and the range of treatment for breast cancer can be greatly improved with early identification [31]. When it comes to early-stage diagnostics, CAD systems powered by AI show a lot of promise.

Automated breast cancer detection utilizing a variety of medical imaging modalities (e.g., thermography, histology, MRI, BUS, CT-scans, etc.) has been the subject of much literature, but there are still many obstacles to overcome. Among the several breast imaging modalities available, "Modality Selection" refers to the process of selecting the one most suited for early-stage diagnosis. To effectively differentiate benign from malignant tumors and enable informed decision-making, tumor discrimination is of utmost importance. Many research relies on manual parameter tuning, which is a time-consuming procedure, to determine the ideal values through experimentation and failed approaches.

Building effective and accurate automated CAD systems relies on having access to large-scale, high-quality datasets, yet publicly available datasets of this kind are in short supply. The reliance on pre-trained models and the paucity of newly-developed models both lead to overfitting and other problems. In addition, previous study has frequently ignored the automated tumor detection phase in favor of statistical approaches and handmade criteria for breast cancer diagnosis; these features typically include color, texture, and morphology.

Rest of the paper is organized as follow: In the next section contribution and objectives of this research are given. In section 2, a detailed literature review is provided. Proposed deep-learning based architecture for breast cancer detection is presented in section 3. In section 4, Experimental results are shown. Finally, a conclusion is given in section 5.

1.1. Contribution

Because most models only know how to label pictures as benign or malignant, they fail to account for the wide variety of breast cancers, including lobular, invasive, and ductal carcinoma in situ. This is a significant shortcoming of current systems that deal with multi-classification. The main goal of this study is to review all the cases where computer-assisted breast cancer screening has benefited from artificial intelligence (AI). We conducted a thorough literature review to identify the challenging factors that affect the accuracy of computer-aided diagnosis in various imaging modalities. A state-of-the-art deep learning method has been created to overcome these challenges in automated breast cancer diagnosis. By delving into topics including modality selection, tumor discrimination, dataset availability, overfitting, scarcity of

new models, handmade features, multi-classification, and manual parameter tweaking, this research hopes to make a significant contribution to the area.

Finally, there is great potential for AI to improve patient outcomes, decrease mistakes, and increase accessibility to healthcare services in healthcare, especially in breast cancer detection. Given these challenges, new approaches are required; this study aims to contribute to this dynamic area by developing a cutting-edge deep learning-based system to aid in the automated early detection of breast cancer.

1.2. Objectives

The goal of this research is to design a deep learning-based computer aided diagnosis system for breast cancer diagnosis. In particular our objective is to work in the following directions.

- Assessing the role of AI in computer aided breast cancer diagnosis: In our study, firstly we have done a comprehensive analysis of applications of AI in automated diagnosis of breast cancer.
- Reviewing Literature: Secondly, we have performed an extensive literature review to identify the challenging parameters (i.e., factors that greatly impact computer aided diagnosis performance). We have considered all imaging modalities during our literature survey.
- Designing a deep model for automated breast cancer diagnosis: A state of the art deep learning-based methodology has been designed for the automated diagnosis of breast cancer.

2. Literature Review

This chapter examined current automated breast cancer diagnosis technologies and their drawbacks. The preceding chapter's assessment parameters are used for a review. We assessed recent computer-aided breast cancer diagnostic classification approaches.

In order to identify breast cancer, mammograms were segmented using dual thresholding in [32]. This method outperforms manual segmentation in terms of speed and memory use. Four sample images with a resolution of 1024 × 1024 pixels are included in the Mini MIAS dataset [43]. Using the input image, the two-stage thresholding algorithm chooses L and U intensity values. A white pixel is located between L and U, whereas all the rest are black. Improve the thresholder image using a 1024 × 1024 mask and morphological border smoothing methods. Restoring the result to the original image reveals any irregularities. To cut down on processing time and expenses for machine learning-based breast tumor identification, another approach [38] proposes an automated image processing technique for segmenting breast lesions in breast MRI images. The method is composed of three steps: first, improving contrast and reducing noise using a median filter; second, using Otsu's thresholding to segment images; and last, using morphological methods to eliminate unwanted regions. The RIDER breast MRI dataset was used for validation, and the results show an accuracy of 97.33% over 150 test photos.

Using breast CT scans, a semi-automated technique for identifying breast lesions is shown in [46]. The three-stage process begins with manually drawing bounding boxes, then uses a four-seed random walk method to partition tumor regions, and last, improves histogram equalization contrast enhancement. A DICE coefficient and a locally acquired dataset consisting of fifty patients are used to validate the proposed method. Using a pseudonymous breast image, [41] demonstrates a basic semi-automated tumor segmentation method. Among the methods employed are sharpening, manual cropping for ROI extraction, global thresholding segmentation, noise reduction, and Sobel edge detection for boundary identification. Shifting gears to more traditional ML models, [17] evaluates SVM performance in binary breast mammography classification using ML and DL characteristics. The work feeds the SVM with size, shape, margin, and intensity for classification after using a region-growing approach to find regions of interest. Based on 607 images, the SVM achieved an area under the curve (AUC) of 0.81. Local Binary Patterns (LBP) for texture feature extraction is suggested in [26] for mammographic image classification into normal and malignant categories. Otsu's thresholding eliminates mammography backgrounds, and the LBP picture

serves as a grid for extracting feature vectors. Using 0.5 as the threshold, SVM classification achieves 84% accuracy on 17,639 DDSM and MIAS images.

Another technique that uses support vector machines to categories mammograms as normal or abnormal is [44]. Using 600 mammography images for training, SVMs with Harris corner detection features achieve 96.8 percent accuracy and 92.5 percent recall on 400 test images. In order to classify breast tumors as either benign or cancerous, a support vector machine (SVM) CAD system is proposed in [20]. Fifty images from the MIAS dataset were cleaned up using median filtering and areas of interest were extracted using split-and-merge segmentation. With a 94% estimation rate, support vector machines (SVMs) employ 13 texture-based characteristics with a grey level co-occurrence matrix (GLCM). A non-linear support vector machine classifier determines if a breast mammography is normal, benign, or malignant in [6]. One hundred and ten Mini-MIAS images make up the training set. Use of a proposed automated pectoral muscle removal method during preprocessing yields 90% correct segmentation features using a Gaussian Mixture Model (GMM). Classification of normal and abnormal breast mammography images using support vector machines is suggested [45]. In preprocessing, masking gets rid of artefacts and maximization gets rid of pectoral muscle. The accuracy rate is 94% when intensity-based features are extracted.

Using K-Nearest Neighbors (KNN) to categories mammographic pictures as normal or abnormal, a computer-aided detection system (CAD) can identify breast cancer at an early stage [15]. Running the algorithm via 120 MIAS pictures yields an accuracy of 92%. [12] employs KNN for binary classification of 110 Mini-MIAS breast mammography as either benign or malignant. A retrieval and normalization of six first-order texture characteristics yields an accuracy of 91.8%. In order to determine if breast micro-calcification is benign or malignant, the CAD system described in [24] employs KNN, SVM, and Decision Tree voting classifiers. For optimal performance, train your top-hat transformation segmentation model on mammograms from the MIAS dataset. One way to detect breast cancer using magnetic resonance imaging (MRI) is outlined in [18]. Using a publicly accessible breast MRI dataset, we can accurately classify normal and abnormal images using a median filter, discrete wavelet feature extraction, and support vector machine classification. Breast tumors may be automatically classified as benign or malignant using a collection of breast MRI images collected locally [34]. The "Relief" feature selection is used to extract and minimize handmade characteristics such as morphological, GLRLM, GLDM, Gabor, and GGLCM. Adaboost fixes data imbalance, which leads to an AUC of 0.9617.

Detailing an automated approach to differentiate between in-situ and infiltrating breast tumors [10]. Radiomics feature extraction, false-positive reduction, and breast area recognition are all accomplished using pre-trained Google Net architecture. An area under the curve (AUC) of 0.70 was achieved on 55 DCE-MRI images using XGBoost classification. In [42], a publicly available thermal imaging dataset is used to validate an automated breast cancer detection algorithm. Some of the preprocessing steps include converting colors to grayscale, reducing noise, improving smoothness, and identifying edges. Using a backpropagation neural network, the extracted features achieve an accuracy of 96.51%. [3] provides a novel strategy to detecting breast cancer by utilizing six methods of texture analysis for feature extraction. The DMR-IR dataset achieves an AUC of 0.989 when a learning-to-rank (LTR) method and multi-layered perceptron are applied.

Automated radiomics features are utilized to categories benign and malignant breast cancers from CT images of the breast [40]. Applying SVM on datasets that are both publicly available and collected locally yields an accuracy rate of 72.54 percent. This is compatible with the recommendations of a recently developed deep Convolutional Neural Network (CNN) model for binary classification that uses extracted features from breast mammography mass images [19]. We use 600 mammographic mass photos from the Digital Database for Screening Mammography (DDSM) to fine-tune and assess the model after training it on ImageNet [11]. We do PCA image whitening after we normalize the recovered images. Upgraded images are subsequently enhanced via data augmentation. High-level and middle-level features are generated using CNN feature extraction. These features are used to train two linear SVM classifiers independently. The testing process involves running both classifiers on an example and prioritizing the results that are consistent. Otherwise, the distance between the test case and the training data designated benign and

malignant examples is calculated and categorized. The proposed model achieved a classification accuracy of 96.7%.

Classifying mammograms as benign or malignant is accomplished using a 7-layered CNN model in an alternative approach [13]. The dataset created by the Medical Image Analysis Society (MIAS) is used for both training and evaluating the models. The dataset contains 95 randomly chosen Regions of Interest (ROIs) for each mammogram, with a 15:4 split between the training and testing datasets. After 500 epochs of CNN model training with the returned ROIs, the highest accuracy was achieved at 0.751. You Only Look Once (YOLO) architecture-based deep convolutional neural network (CNN) for breast tumor identification and malignant/benign classification is also available in [37]. In order to train the model, DDSM mammograms are used. Images are normalized and noise and background are removed during preprocessing. Random rotation is one example of the data augmentation applied to images. Bounding boxes are used to extract and resize mass patches. Prior to training all mammography patches, mass patches are taught. Main training uses initial training weights to initialize modified CNN weights. The proposed system achieves a detection accuracy of 90% and a classification accuracy of 93.5%. For breast mammography classification, [9] presents a 9-layered CNN trained on MIAS. Images from the training dataset are used to conduct morphological closure and masking in order to segment ROIs. The suggested model, when fed resized images, obtains a multi-classification and binary accuracy of 65%.

An additional CNN is presented in [28] for the purpose of binary breast mammography classification. 116 MIAS anomalous photos are used to train. Image preprocessing includes global contrast normalization and cropping to eliminate noise and backgrounds. The next stage is to eliminate ROIs and replenish the data to minimize overfitting and achieve dataset balance. By feeding data into the CNN, we can see how the model's generalizability is affected by the CNN's depth and dropout adaptation. The results show that deep CNN classification is improved when a dropout approach and a smaller filter size are used. This study details a web-based CAD system that uses deep convolutional neural networks (CNNs) to binary classify breast histopathology images [7]. To teach the model, we drew from BreaKHis, an open-source database of high-resolution breast images. The input images to a seven-layer deep convolutional neural network (CNN) are 192 × 192 pixels in size. With 6327 examples trained and 1582 cases tested, the model achieved a 99% training accuracy and a 91.4% testing accuracy. [39] demonstrates a U-Net-inspired deep encoder-decoder CNN design for mammography-based early breast cancer detection. Microcalcification and bulk detection subsets of the CBIS-DDSM database are used to train the model. The training and assessment phases involved retrieving 692 masses and 603 microcalcified people from the database. We equalize the contrast of the histogram and randomly rotate it as part of the preprocessing.

The proposed CNN achieves an accuracy of 95.01% when classifying masses and 94.31% while classifying micro-calcifications. Uses both subjective and objective criteria to classify mammograms [47]. A proprietary dataset consisting of 400 mammography images is used to train the algorithm. Adaptive mean filter is used to reduce noise and increase the mass-environment contrast. Partitioning the ROIs into 48x48 subregions, bounding boxes are used to extract them. A feature vector of 24 lengths is generated from these subregions by use of a 7-layer convolutional neural network (CNN). Using the recovered attributes, an Unsupervised Extreme Learning Machine (US-ELM) technique may cluster areas that are suspicious and those that are not. The proposed method employs an ELM classifier for the classification of five morphological, five textural, and seven density-based attributes. When it comes to classification using all four characteristics, experimental results show that ELM is superior to SVM.

For the purpose of automatically distinguishing between benign and cancerous breast thermograms, [35] also details a deep learning architecture. The author uses the publicly available breast thermal imaging DMR-IR dataset for training and validation purposes. Normalization, scaling, and RGB-to-gray conversion are all part of the preprocessing. This proposed network uses 5 2D-convolution, 3 max-pooling, 2 fully connected, batch normalization, and dropout layers to get an accuracy of 95.8%. Three ResNet-34/50/101 ensemble classifiers are compared in this study: one that trains from scratch, one that fine-tunes the whole learned model, and one that tweaks only the last layer. Without the use of pre-trained weights, the suggested ensemble classifier is trained on the BACH-2018 dataset of microscopic breast images. We employ data

augmentation and image patching. Validation accuracy for the model is 97.3% and test accuracy is 86%. Not only that, but [14] evaluates several Transfer Learning (TL) approaches to binary categorization of breast histology WSI patches. Starting with 5,000 patches extracted from the locally collected WSIs dataset, the AlexNet and GoogLeNet architectures undergo a full training cycle. Augmenting data increases the size of datasets. Compared to GoogLeNet's 94.12% final accuracy, AlexNet's is 93.40%. In [27], the author employs an IRRCNN to identify binary and multi-class breast cancer images from histology slides. To put the model to the test, data is supplemented to the BreakHis and BioImaging-2015 datasets. By integrating residual networks with recurrent convolutional neural networks, the proposed architecture enhances performance.

Using a specially developed mammography dataset, [17] evaluates three support vector machine classifiers for the subjective and objective classification of benign and malignant breast lesions. Before normalization, mammograms are down sampled to 256 × 256 and divided into 512 × 512 patches. Each of the three classifiers undergoes a mixed-methods training process that includes deep learning, handcrafted features, and a pre-trained convolutional neural network (CNN) called AlexNet [27]. Because of its low dimensionality and good performance, Fc6 is used in feature selection by layer-wise support vector machine classifiers. [8] uses six features to categories BACH-2018 microscopic images. The feature extraction process makes use of five pretrained deep CNNs from ImageNet: ResNet-18, ResNet-152, ResNeXt, NASNet-A, and VGG16, as well as two new features, PFTAS and GLCM. In order to get an accuracy of 79% on Part A of the BACH challenge, the features are processed using Multi-View Random Forest Kernel SVM (RFSVM). A radiomics model separates benign from malignant breast tumors utilizing shape and GLGM, in addition to deep learning (AlexNet) [30]. With 212 unique digital mammography images, the model achieves a 96.4% success rate. In [14], 121 cases are used to construct a hybrid approach for classifying histopathological full slide photos. After removing noise and extracting patches, CNN obtains enhanced data. Different approaches to transfer learning are investigated for AlexNet and GoogLeNet, both with and without pretrained weights. For the purpose of picture texture characterization, GLCM and LBP features are recovered, and SVM trained over them attains a 98.90% accuracy rate. [49] evaluates a pretrained VGG16 CNN and a hybrid model (VGG-16 + SVM) for multi-classification accuracy on breast histopathology photos. With a validation accuracy of 81.25% and a Bioimaging 2015 accuracy of 80.6%, the hybrid approach fine-tunes a pretrained VGG-16 utilizing BACH-2018. There is a hybrid approach to classifying mammograms for the purpose of detecting breast cancer [4]. It employs a pretrained convolutional neural network (CNN) based on transfer learning (VGG-16) and a set of linear classifiers. The MIAS dataset is used to enhance classification accuracy with linear classifiers such as KNN, Decision Tree, and Gradient Boosting. One method for classifying mass patches found in many mammograms is the MV-DNN Deep Neural Network, which is shown in [48]. Thanks to its 0.89 AUC and 85% accuracy, MV-DNN shines on the BCDR-F03 dataset.

To locate and categories breast lumps, a Faster Recurrent CNN employs five feature extraction models [53]. With a typical accuracy of 0.85, Inception ResNet-V2 outperforms MobileNet, which has the lowest at 0.60. A recent study employed the DeepLab-V3 and Mask-RCNN frameworks to semantically segment and categories breast masses [5]. Applying Mask-RCNN and DeepLab-V3, we achieve 98% and 95% accuracy, respectively, using both MIAS and CBIS-DDSM datasets. Automating BUS ROI extraction is achieved by combining Faster RCNN with Inception-ResNet-v2 for object recognition [50]. Significant improvement in intersection over union (IOU) is demonstrated with the approach on two BUS image datasets. takes a look at five distinct methods to binary breast tumor classification using multi-parametric MRI [16]. With the help of a pretrained VGG-19 CNN, features are extracted, and SVM carries out the classification. We look into methods that combine features, images, and classifiers, and we find that they work. [33] proposes automatically classifying breast thermograms using pretrained Inception-V3 architecture. It takes 15 epochs for SVM to classify the feature vector accurately if the probability of a cancerous zone is 0.5 to 0.6. In order to classify breast images from tomosynthesis and 2D and 3D mammography, [52] examines eleven different deep learning architectures. ROCs of 0.7274 and 0.6632 for

2D and 3D photos, respectively, show that AlexNet's transfer learning outperforms other networks in categorization.

3. PROPOSED ARCHITECTURE

Our study's recommended architecture is shown below:

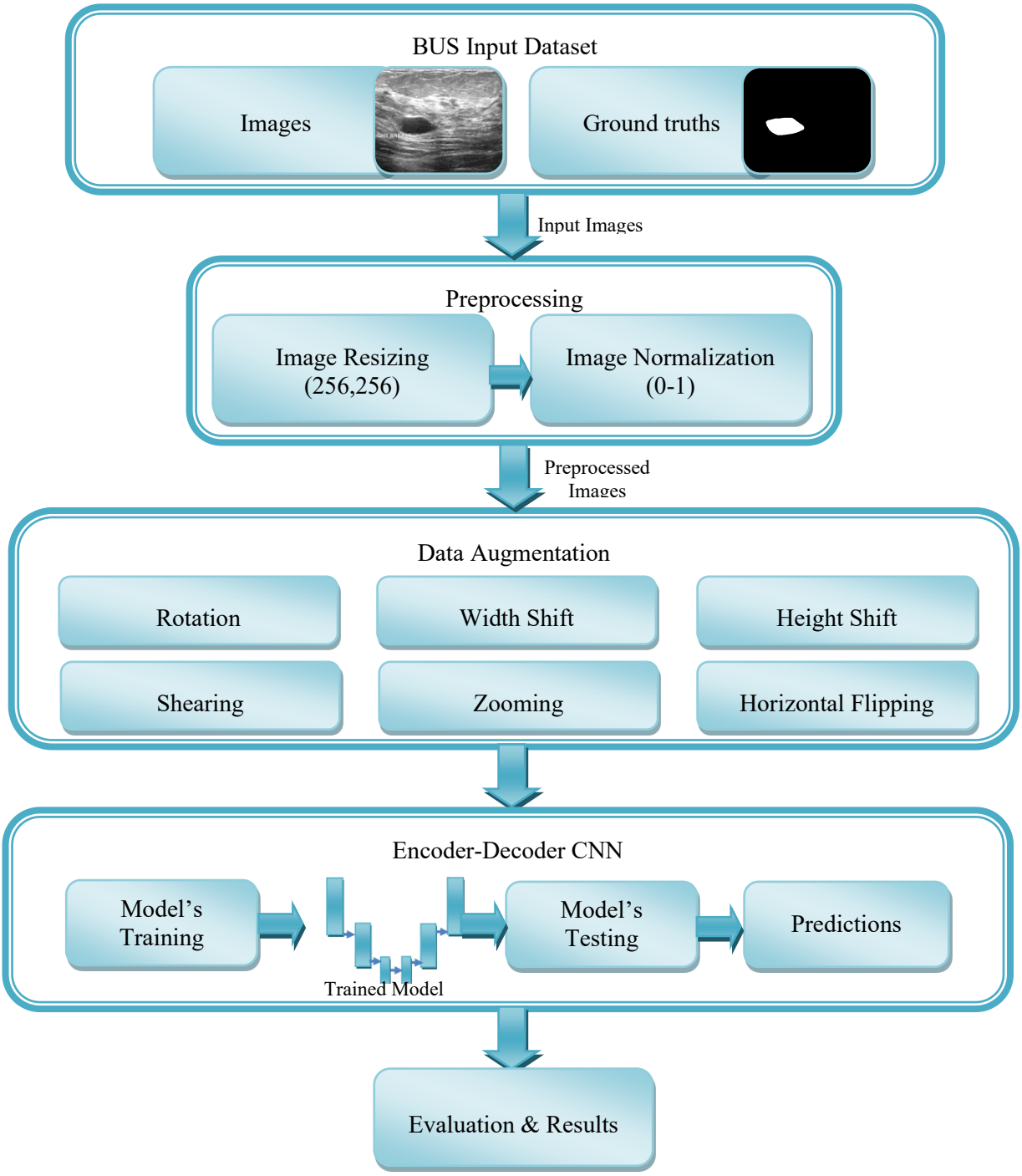

Figure 1: Proposed Architecture

3.1. Datasets Description

3.1.1. Dataset 1

In this investigation, we used a publicly accessible breast ultrasound imaging dataset [29] of 780 ultrasound pictures for training and evaluating the suggested deep learning architecture. This dataset was gathered in 2018 at Baheya Hospital in Egypt. The data was originally collected in 1280*1024 resolution DICOM format and subsequently transformed to 500*500 pixels average PNG size. The three main categories of images are benign, normal, and malignant.

Table 1: Training dataset description

Case	No of Images
Malignant	210
Benign	487
Normal	133
Total	780

3.1.2. Dataset 2

We also used a publicly accessible BUS imaging dataset, UDIAT [51], to evaluate trained models, which contains 163 BUS pictures (110 benign & 53 malignant) in PNG format with a size of 760*570 pixels.

3.2. Preprocessing

Dataset images are first preprocessed before being fed into the proposed deep architecture. During the preprocessing stage, two tasks are completed:

3.2.1 Image Resizing

All of the images in the collection are of different sizes, but on average they have a 500*500 pixel resolution. In contrast, deep convolutional neural network design strictly enforces the use of identically sized input images. To address this issue, two potential options are as follows: The first is padding, which can be either zero, the same, or constant. image resizing by the inclusion of extra columns and rows, which might result in higher processing expenses. 2) Resizing photos: Reducing the computational cost involves resizing photos to make their sizes equal. In this example, we resized the input photographs to 256*256 pixels so that they would all be the same size.

3.2.2. Image Normalization

Starting deep model weights with small, arbitrary values between 0 and 1 and adjusting them according to the loss calculated through an optimization process is a common practice. A wide intensity range (i.e., 0-255) is typically present in input images and ground facts. In the absence of intensity normalization or rescaling, problems such as an inflated gradient could arise from using such images, rendering the learning process useless. To avoid these kinds of issues, we rescale input photos and ground facts, also known as 0-1 normalization, before delivering them to the network.

3.3. Data Augmentation

Over-fitting is a key issue that arises during the training of deep learning architectures (i.e., the model memorizes input data and produces better results during training than testing outcomes). One of the primary reasons of this problem is model training on a little amount of data. Due to the scarcity of large-scale medical imaging datasets, this problem frequently occurs during automated illness diagnosis. To prevent such issues, a data augmentation approach is used (i.e., multiple transformations are done to input photos to increase the size of the dataset). We used six distinct forms of transformation of input data in our suggested technique, including:

- Rotation: photos are rotated at random in a range of 0.2 degrees.
- Width Shift: In this method, input photos are randomly horizontally moved to the left or right by a factor of 0.05% of their overall width.
- Height Shift: In this method, incoming photos are randomly vertically moved higher or lower by 0.05% of their entire height.
- Shearing: the input pictures are sheared counterclockwise in a 0.05-degree range.
- Zoom: input photos are randomly zoomed in the range [1-0.05, 1+0.05].
- Horizontal Flipping: This method randomly flips input photos horizontally.

3.4. Model's Architecture

Using our suggested semantic segmentation architecture as a benchmark, we evaluated its performance against that of three leading encoder-decoder semantic segmentation CNN designs: UNet [25], UNet++, and SegNet [36]. We will quickly go over these three state-of-the-art semantic segmentation architectures.

3.5. Existing Semantic Architectures

3.5.1. UNet

UNet design is divided into two sections: a contracting road (encoder) and a costly path (decoder). The encoder section comprises of two successive 3*3 convolution layers, followed by a ReLU and 2*2 Max pooling layers for picture downsampling, with the number of convolutional filters increased (from 64 to 128) for each downsampling layer.

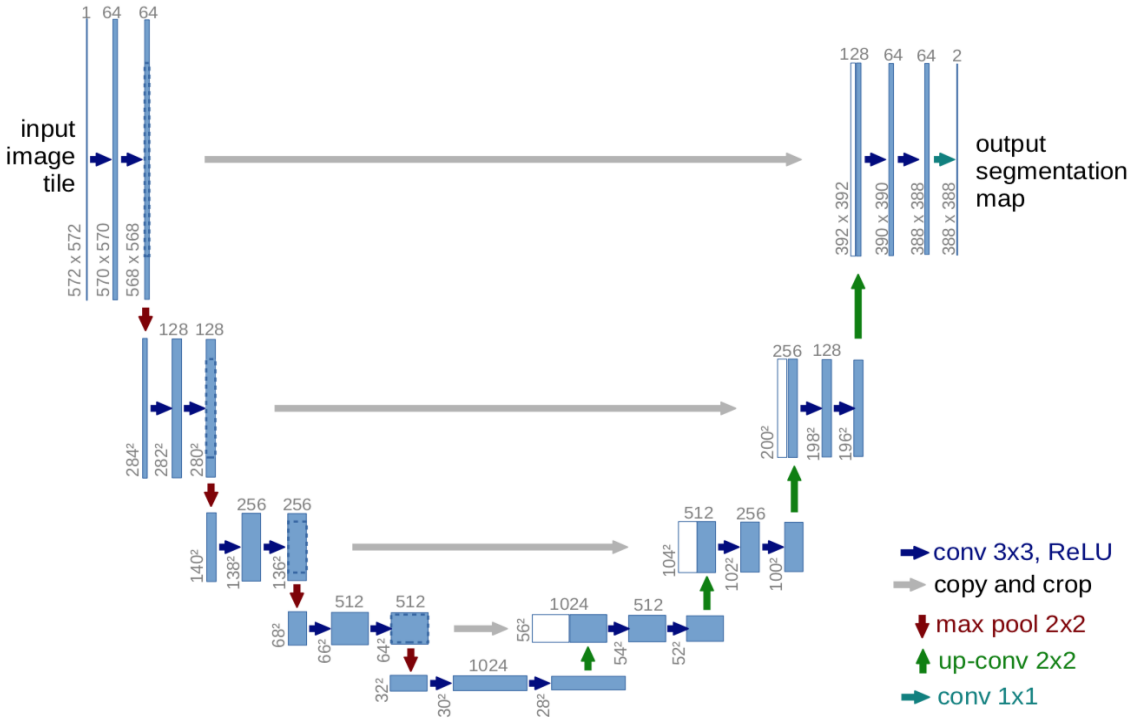

Figure 2: UNet Architecture

The decoder or expanding section includes feature modules for up-sampling, such as a 2*2 convolutional layer (up-convolution) that reduces the amount of feature maps by half (from 128 to 64) and a concatenation layer. Finally, 1*1 convolutional layer is integrated at the final layer to map 64*64 feature mappings to the necessary number of classes. The graphic above depicts the basic UNet architecture.

3.5.2. UNet++

The currently available UNet++ is built on top of the UNet architecture, and it uses dense convolutional blocks along each skip link to gather each previous feature map. UNet++ achieves this by adjusting the skip connections between the encoder and decoder. There are three main ways in which UNet++ differs from UNet: 1) A deep supervision mechanism is used; 2) dense skip connections are used to improve gradient flow; and 3) encoder and decoder feature maps are connected using convolutional blocks to bridge the semantic gap. The following diagram depicts the fundamental architecture of UNet++.

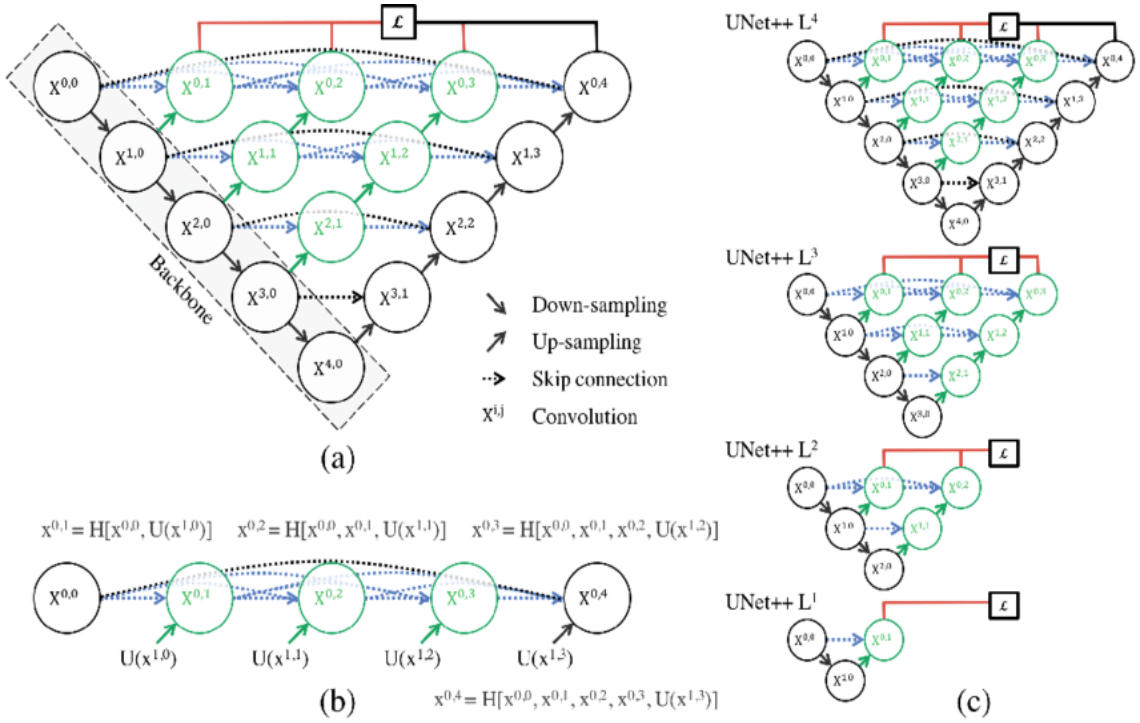

Figure 3: UNet++ Architecture

3.5.3. SegNet

The state-of-the-art SegNet architecture is another unique fully convolutional neural network whose performance was examined in our study. This design is also divided into two parts (encoder and decoder). The architecture of the encoder part is similar to a 13-layer convolutional neural network (CNN), more especially VGG-16, with 13 layers in the decoder part for every layer of the encoder. In each encoder block, you'll find convolutional layers, batch normalization, ReLU, and max pooling layers; in each decoder block, you'll find an up-sampling layer in place of the pooling layer. SegNet's encoder and decoder components do not share any skip links with UNet or UNet++. The picture below shows the SegNet architecture.

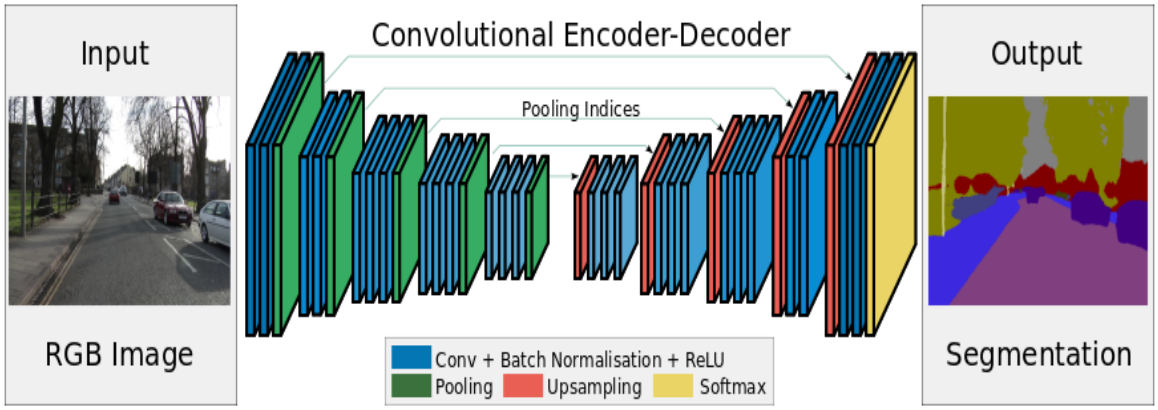

Figure 4: SegNet Architecture

3.5.4. Proposed Semantic Architecture

The proposed design is an upgraded variant of the state-of-the-art SegNet architecture. What sets SegNet apart from SegNet-xd is the inclusion of skip connections in the updated proposed architecture. The newly proposed model is simpler and uses less computing power than conventional SegNet since SegNet-xd has fewer convolutional layers. There are two kinds of skip connections in the updated model: 1) between the encoder's convolutional layers, and 2) between the decoder's and encoder's corresponding convolutional layers. Batch normalization also happens after every convolutional layer in SegNet, while in SegNet-xd it happens only after the second convolutional layer in each encoder and decoder block. Below is a diagram demonstrating the proposed SegNet-xd architecture.

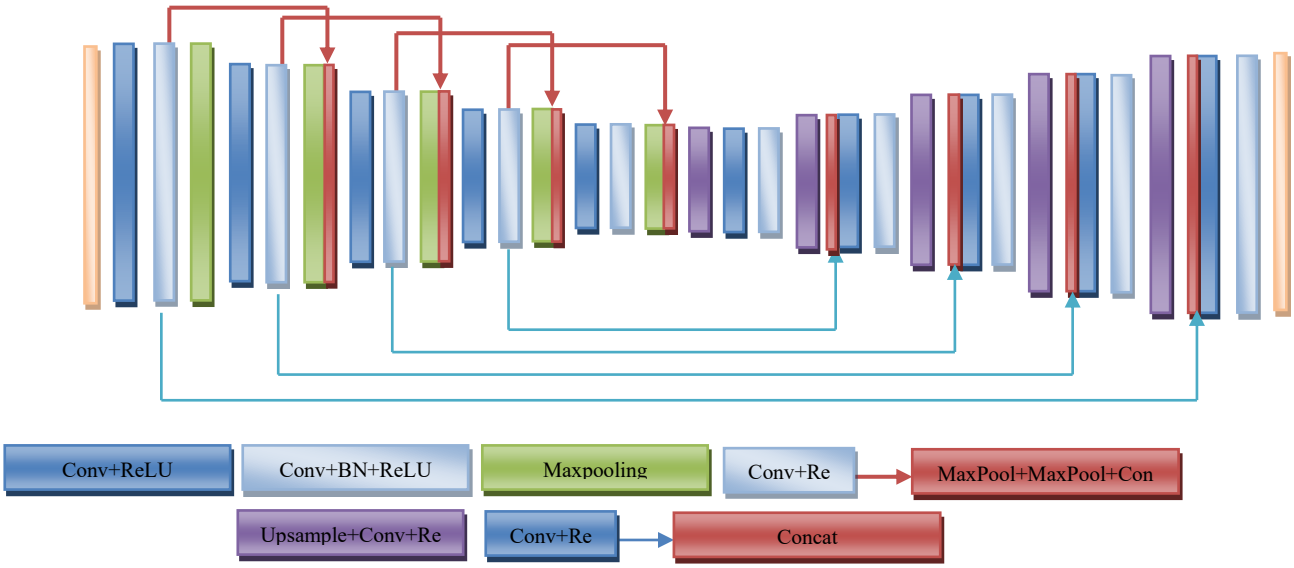

Figure 5: Proposed Semantic Segmentation Architecture

3.6. Training Parameters

The proposed deep CNN is trained over 500 epochs, utilizing 70% of the training data and a batch size of 10. Three sections of Dataset1 70% for training, 20% for validation, and 10% for testing, are utilized to

train the proposed model. We keep track of validation loss and the best model's weights throughout training. A binary cross-entropy loss and a learning rate of lr = 1e-4 were employed by the Adam optimizer.

4. Results and evaluation

In this chapter, we examined the outcomes of a suggested deep CNN architecture for automated breast lesion segmentation. The following are the outcomes of the experiments:

4.1. Experimental Results

Using 546 images for testing and 156 images for validation, with an equal percentage of benign, malignant, and normal cases, the newly proposed semantic segmentation architecture and the three leading ones are trained over 500 epochs. The training and validation loss of distinct architectures are shown in the following graphs.

4.1.1 Training and Validation Loss Plots

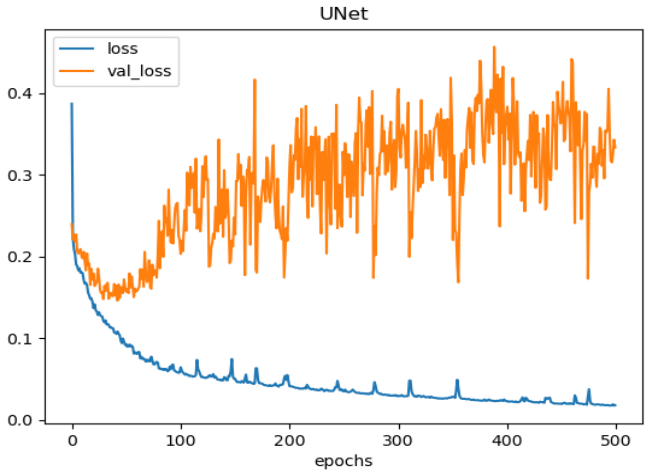

Figure 6: Training and Validation Loss of UNet Architecture

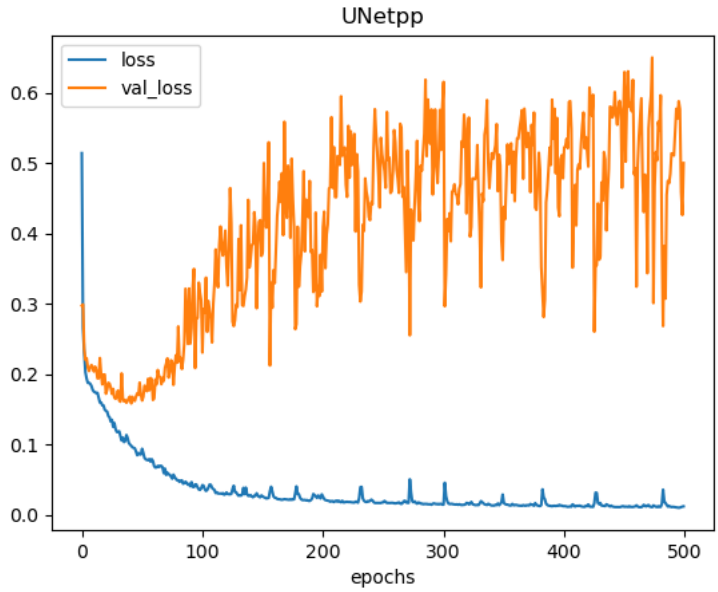

Figure 7: Training and Validation Loss of UNet++ Architecture

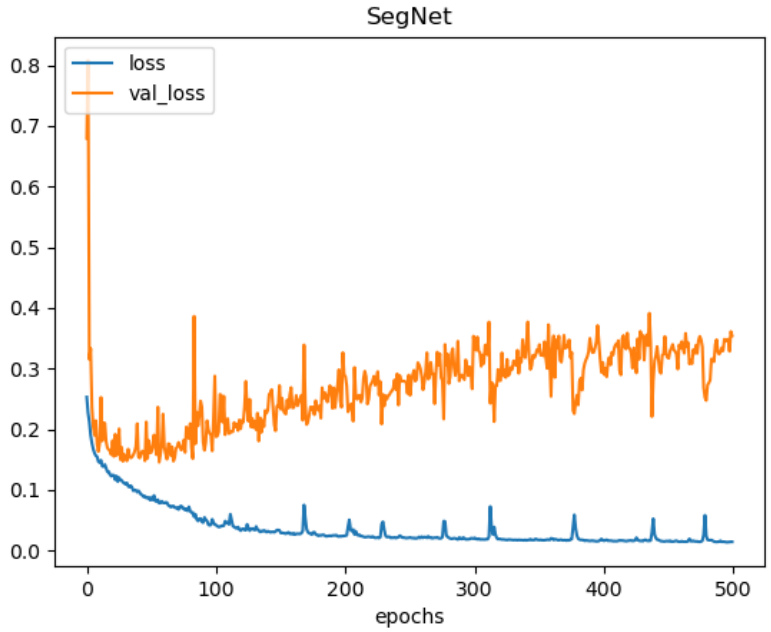

Figure 8: SegNet Architecture Training and Validation Loss

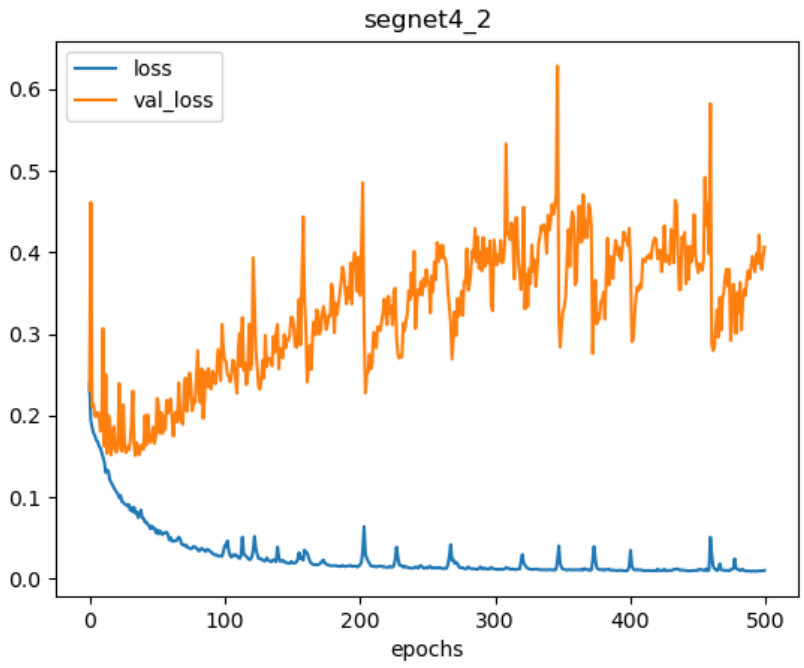

Figure 9: Proposed Architecture Training and Validation Loss

4.1.2. Model's Predictions

The following figures display the forecasts of the current architecture, two test photographs, and the original mask images, as well as the suggested design and three other architectures:

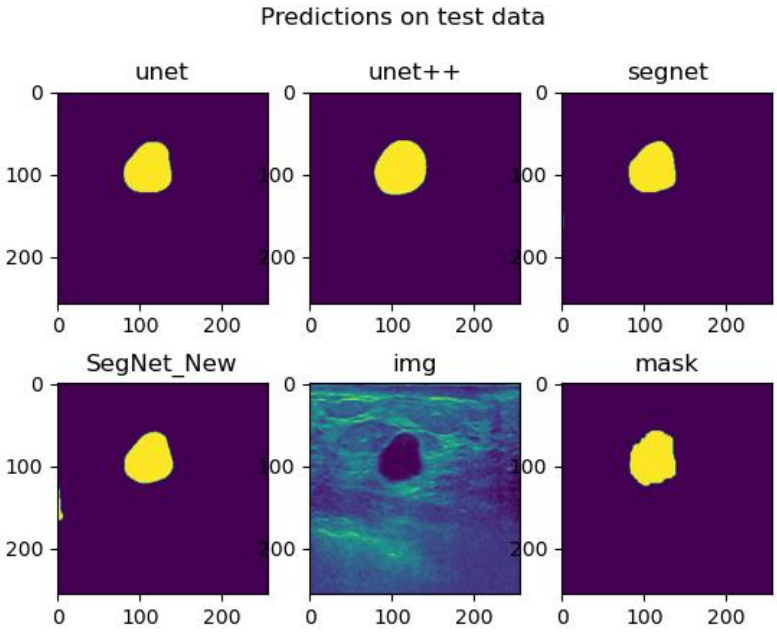

Figure 10: Predictions over 1st testing image

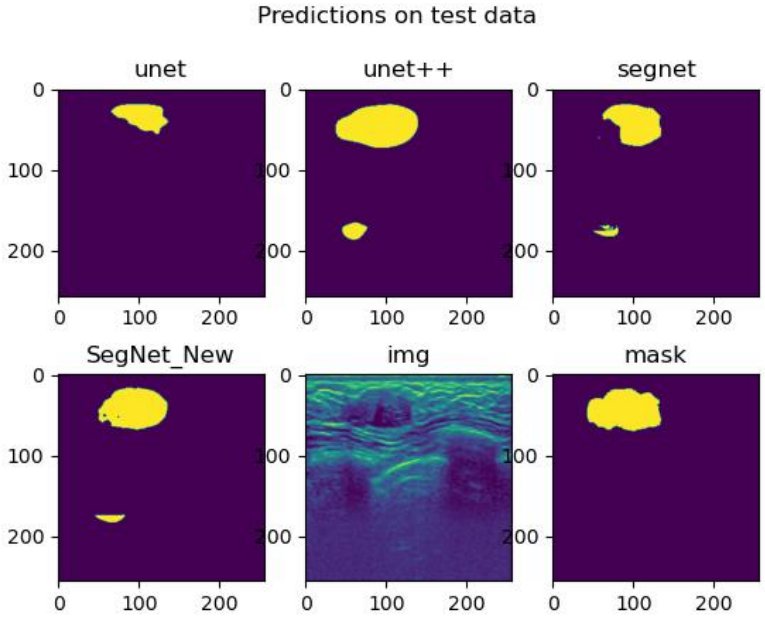

Figure 11: Predictions over 2nd testing image

4.1.3. Jaccard Scores over Dataset1

Table 2: Dataset1 Jaccard Score

	0.5	**0.4**	**0.3**	**0.1**	**0.05**
UNet	0.740	0.743	0.746	0.752	0.753
UNet++	0.722	0.726	0.729	0.734	0.734
SegNet	0.745	0.748	0.750	0.750	0.750
Proposed Architecture	*0.761*	*0.763*	*0.765*	*0.769*	*0.770*

4.1.4. Jaccard Scores over Dataset2

Table 3: Dataset2 Jaccard Score

	0.5	**0.4**	**0.3**	**0.2**	**0.1**	**0.01**
UNet	0.694	0.696	0.698	0.700	0.703	0.708
UNet++	0.715	0.714	0.714	0.712	0.709	0.697
SegNet	*0.774*	*0.777*	*0.775*	*0.770*	*0.762*	0.725
Proposed Architecture	0.744	0.744	0.745	0.746	0.747	*0.748*

4.2. Discussion:

To compare the four deep segmentation architectures' predictions from Datasets 1 and 2, we may see their Jaccard Scores in the tables up top. Determine the Jaccard score by thresholding the final prediction from each of the four architectures at different levels, in order to look at genuine positives and false positives. Our new design outperformed the three previous designs using Dataset 1 to the tune of 0.77 on the Jaccard scale. In contrast, our proposed architecture achieved second-best results on Dataset2, whereas SegNet generated the best predictions (i.e., from 0.1 to 0.5), albeit with a higher number of false positives. The 0.01 cutoff was where our suggested design really shone. The proposed architecture outperforms UNet++ and SegNet in terms of computational efficiency without sacrificing performance.

5. Conclusion

The application of artificial intelligence in medicine has transformed it. It is difficult to detect some fatal diseases by manually examining and properly interpreting numerous medical datasets (e.g., medical pictures, medical reports). Manual diagnosis is fraught with subjectivity and misinterpretation. Because of their similarity in characteristics, precise interpretation and distinction of tumors (i.e. benign and malignant) by manual study is a difficult task. Deep learning may be used to overcome these problems because to its ability to learn thousands of high-level and low-level features. A lot of work has been done in recent years to automate the identification of breast cancer using breast imaging datasets from several modalities. The majority of the work, however, has been done with X-Ray mammogram datasets. The absence of large-scale publicly available quality datasets is a critical concern while developing such systems. The bulk of publicly available datasets are either tiny in size or have a lot of noise in them. We used a recently available breast ultrasound imaging dataset to detect aberrations in this study. A unique deep CNN was presented for semantic segmentation of these images. The model is trained on 70% of the available data, confirmed on 20%, and tested on 10%. Furthermore, another publicly available small size dataset was utilized to assess the proposed approach. Our proposed model earned a Jaccard score of 0.77 when tested on Dataset1 and 0.748 when tested on Dataset2. We want to improve the outcomes of our proposed model in the future by using alternative regularization procedures, as the model performed better on training data than on testing data.

References

[1] Yu, Kun-Hsing, Andrew L. Beam, and Isaac S. Kohane. "Artificial intelligence in healthcare." *Nature biomedical engineering* 2, no. 10 (2018): 719-731.

[2] Scott, Lisa. "Medication errors... This reflective account is based on NS776 Cloete L (2015) Reducing medication errors in nursing practice. Nursing Standard. 29, 20, 50-59." *Nursing Standard* 30, no. 35 (2016).

[3] Abdel-Nasser, Mohamed, Antonio Moreno, and Domenec Puig. "Breast cancer detection in thermal infrared images using representation learning and texture analysis methods." *Electronics* 8, no. 1 (2019): 100.

[4] Agrawal, Sanket, Rucha Rangnekar, Divye Gala, Sheryl Paul, and Dhananjay Kalbande. "Detection of breast cancer from mammograms using a hybrid approach of deep learning and linear classification." In *2018 International Conference on Smart City and Emerging Technology (ICSCET)*, pp. 1-6. IEEE, 2018.

[5] Ahmed, L., M. M. Iqbal, H. Aldabbas, S. Khalid, Y. Saleem, and S. Saeed. "Images data practices for semantic segmentation of breast cancer using deep neural network. J Ambient Intell Humaniz Comput." (2020).

[6] Arafa, Amany Abdel Aziz, Nesma El-Sokary, Ahmed Asad, and Hesham Hefny. "Computer-aided detection system for breast cancer based on GMM and SVM." *Arab Journal of Nuclear Sciences and Applications* 52, no. 2 (2019): 142-150.

[7] Arslan, Ahmet Kadir, Şeyma Yaşar, and Cemil Çolak. "Breast cancer classification using a constructed convolutional neural network on the basis of the histopathological images by an interactive web-based interface." In *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 1-5. IEEE, 2019.

[8] Cao, Hongliu, Simon Bernard, Laurent Heutte, and Robert Sabourin. "Dissimilarity-based representation for radiomics applications." *arXiv preprint arXiv:1803.04460* (2018).

[9] Charan, Saira, Muhammad Jaleed Khan, and Khurram Khurshid. "Breast cancer detection in mammograms using convolutional neural network." In *2018 international conference on computing, mathematics and engineering technologies (iCoMET)*, pp. 1-5. IEEE, 2018.

[10] Conte, Luana, Benedetta Tafuri, Maurizio Portaluri, Alessandro Galiano, Eleonora Maggiulli, and Giorgio De Nunzio. "Breast Cancer Mass Detection in DCE–MRI Using Deep-Learning Features Followed by Discrimination of Infiltrative vs. In Situ Carcinoma through a Machine-Learning Approach." *Applied Sciences* 10, no. 17 (2020): 6109.

[11] Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248-255. Ieee, 2009.

[12] Diaz, Ricky Aurelius Nurtanto, Ni Nyoman Tria Swandewi, and Kadek Dwi Pradnyani Novianti. "Malignancy determination breast cancer based on mammogram image with k-nearest neighbor." In *2019 1st International Conference on Cybernetics and Intelligent System (ICORIS)*, vol. 1, pp. 233-237. IEEE, 2019.

[13] Guan, Shuyue, and Murray Loew. "Breast cancer detection using transfer learning in convolutional neural networks." In *2017 IEEE applied imagery pattern recognition workshop (AIPR)*, pp. 1-8. IEEE, 2017.

[14] He, Simin, Jun Ruan, Yi Long, Jianlian Wang, Chenchen Wu, Guanglu Ye, Jingfan Zhou, Junqiu Yue, and Yanggeling Zhang. "Combining deep learning with traditional features for classification and segmentation of pathological images of breast cancer." In *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 1, pp. 3-6. IEEE, 2018.

[15] Htay, Than Than, and Su Su Maung. "Early stage breast cancer detection system using glcm feature extraction and k-nearest neighbor (k-NN) on mammography image." In *2018 18th International Symposium on Communications and Information Technologies (ISCIT)*, pp. 171-175. IEEE, 2018.

[16] Hu, Qiyuan, Heather M. Whitney, and Maryellen L. Giger. "A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI." *Scientific reports* 10, no. 1 (2020): 10536.

[17] Huynh, Benjamin Q., Hui Li, and Maryellen L. Giger. "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks." *Journal of Medical Imaging* 3, no. 3 (2016): 034501-034501.

[18] Ibraheem, Amira Mofreh, Kamel Hussein Rahouma, and Hesham FA Hamed. "Automatic MRI breast tumor detection using discrete wavelet transform and support vector machines." In *2019 Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, vol. 1, pp. 88-91. IEEE, 2019.

[19] Jiao, Zhicheng, Xinbo Gao, Ying Wang, and Jie Li. "A deep feature based framework for breast masses classification." *Neurocomputing* 197 (2016): 221-231.

[20] Jothilakshmi, G. R., and Arun Raaza. "Effective detection of mass abnormalities and its classification using multi-SVM classifier with digital mammogram images." In *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, pp. 1-6. IEEE, 2017.

[21] Khan, Muhammad Owais. "The necessity of awareness of breast cancer amongst women in Pakistan." *JPMA. The Journal of the Pakistan Medical Association* 59, no. 11 (2009): 804.

[22] Nayab Khan, Riaz Ahmad, Muhammad Nadeem, and Iftikhar Hussain. "Influence of Education and Socio-Economic Factors on Stage of Cancer Diagnosis: A Study in Pakistani Population." *Annals of PIMS ISSN* 1815: 2287.

[23] Khokher, Samina, Muhammad Usman Qureshi, Saqib Mahmood, and Sadia Sadiq. "Determinants of advanced stage at initial diagnosis of breast cancer in Pakistan: adverse tumor biology vs delay in diagnosis." *Asian Pacific Journal of Cancer Prevention* 17, no. 2 (2016): 759-765.

[24] Khoulqi, Ichrak, and Najlae Idrissi. "Breast cancer image segmentation and classification." In *Proceedings of the 4th International Conference on Smart City Applications*, pp. 1-9. 2019.

[25] Kook, Shin-Ho, Hae-Won Park, Young-Rae Lee, Young-Uk Lee, Won-Kil Pae, and Yong-Lai Park. "Evaluation of solid breast lesions with power Doppler sonography." *Journal of clinical ultrasound* 27, no. 5 (1999): 231-237.

[26] Kral, Pavel, and Ladislav Lenc. "LBP features for breast cancer detection." In *2016 IEEE international conference on image processing (ICIP)*, pp. 2643-2647. IEEE, 2016.

[27] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks." *Communications of the ACM* 60, no. 6 (2017): 84-90.

[28] Li, Bin, Yunhao Ge, Yanzheng Zhao, Enguang Guan, and Weixin Yan. "Benign and malignant mammographic image classification based on convolutional neural networks." In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, pp. 247-251. 2018.

[29] Al-Dhabyani, Walid, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. "Dataset of breast ultrasound images." *Data in brief* 28 (2020): 104863.

[30] Liang, Ming, and Xiaolin Hu. "Recurrent convolutional neural network for object recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3367-3375. 2015.

[31] Badawy, Samir M., Alaa A. Hefnawy, Hassan E. Zidan, and Mohammed T. GadAllah. "Breast cancer detection with mammogram segmentation: a qualitative study." *International Journal of Advanced Computer Science and Applications* 8, no. 10 (2017).

[32] Magny, Samuel J., Rachel Shikhman, and Ana L. Keppke. "Breast imaging reporting and data system." In *StatPearls [Internet]*. StatPearls publishing, 2022.

[33] Marami, Bahram, Marcel Prastawa, Monica Chan, Michael Donovan, Gerardo Fernandez, and Jack Zeineh. "Ensemble network for region identification in breast histopathology slides." In *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15*, pp. 861-868. Springer International Publishing, 2018.

[34] Lu, W. L., L. Jansen, W. J. Post, J. Bonnema, J. C. Van de Velde, and G. H. De Bock. "Impact on survival of early detection of isolated breast recurrences after the primary treatment for breast cancer: a meta-analysis." *Breast cancer research and treatment* 114 (2009): 403-412.

[35] Mohebian, Mohammad R., Hamid R. Marateb, Marjan Mansourian, Miguel Angel Mañanas, and Fariborz Mokarian. "A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning." *Computational and structural biotechnology journal* 15 (2017): 75-85.

[36] Pavan, Ana LM, Antoine Vacavant, Allan FF Alves, Andre P. Trindade, and Diana R. de Pina. "Automatic identification and extraction of pectoral muscle in digital mammography." In *World Congress on Medical Physics and Biomedical Engineering 2018: June 3-8, 2018, Prague, Czech Republic (Vol. 1)*, pp. 151-154. Springer Singapore, 2019.

[37] Ponraj, D. Narain, M. Evangelin Jenifer, P. Poongodi, and J. Samuel Manoharan. "A survey on the preprocessing techniques of mammogram for the detection of breast cancer." *Journal of Emerging Trends in Computing and Information Sciences* 2, no. 12 (2011): 656-664.

[38] Ramani, R., S. Suthanthiravanitha, and S. Valarmathy. "A survey of current image segmentation techniques for detection of breast cancer." *International Journal of Engineering Research and Applications (IJERA)* 2, no. 5 (2012): 1124-1129.

[39] Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779-788. 2016.

[40] Saphner, Thomas, Douglass C. Tormey, and Robert Gray. "Annual hazard rates of recurrence for breast cancer after primary therapy." *Journal of clinical oncology* 14, no. 10 (1996): 2738-2746.

[41] Smirnov, Evgeny A., Denis M. Timoshenko, and Serge N. Andrianov. "Comparison of regularization methods for imagenet classification with deep convolutional neural networks." *Aasri Procedia* 6 (2014): 89-94.

[42] Song, Enmin, Luan Jiang, Renchao Jin, Lin Zhang, Yuan Yuan, and Qiang Li. "Breast mass segmentation in mammography using plane fitting and dynamic programming." *Academic radiology* 16, no. 7 (2009): 826-835.

[43] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems* 27 (2014).

[44] Tahir, Bilal, Sajid Iqbal, M. Usman Ghani Khan, Tanzila Saba, Zahid Mehmood, Adeel Anjum, and Toqeer Mahmood. "Feature enhancement framework for brain tumor segmentation and classification." *Microscopy research and technique* 82, no. 6 (2019): 803-811.

[45] Vogel, Victor G. "Assessing risk of breast cancer: Tools for evaluating a patient's 5-year and lifetime probabilities." *Postgraduate Medicine* 105, no. 6 (1999): 49-58.

[46] Wang, Hongyu, Jun Feng, Zizhao Zhang, Hai Su, Lei Cui, Hua He, and Li Liu. "Breast mass classification via deeply integrating the contextual information from multi-view data." *Pattern Recognition* 80 (2018): 42-52.

[47] Wang, Na, Cheng Bian, Yi Wang, Min Xu, Chenchen Qin, Xin Yang, Tianfu Wang, Anhua Li, Dinggang Shen, and Dong Ni. "Densely deep supervised networks with threshold loss for cancer detection in automated breast ultrasound." In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*, pp. 641-648. Springer International Publishing, 2018.

[48] Wang, Yaqi, Lingling Sun, Kaiqiang Ma, and Jiannan Fang. "Breast cancer microscope image classification based on CNN with image deformation." In *Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15*, pp. 845-852. Springer International Publishing, 2018.

[49] Wang, Yi, Chenchen Qin, Chuanlu Lin, Di Lin, Min Xu, Xiao Luo, Tianfu Wang, Anhua Li, and Dong Ni. "3D Inception U-net with asymmetric loss for cancer detection in automated breast ultrasound." *Medical Physics* 47, no. 11 (2020): 5582-5591.

[50] Yap, Moi Hoon, Gerard Pons, Joan Marti, Sergi Ganau, Melcior Sentis, Reyer Zwiggelaar, Adrian K. Davison, and Robert Marti. "Automated breast ultrasound lesions detection using convolutional neural networks." *IEEE journal of biomedical and health informatics* 22, no. 4 (2017): 1218-1226.

[51] Yemini, Mor, Yaniv Zigel, and Dror Lederman. "Detecting masses in mammograms using convolutional neural networks and transfer learning." In *2018 ieee international conference on the science of electrical engineering in israel (icsee)*, pp. 1-4. IEEE, 2018.

[52] Zhang, Yungang, Bailing Zhang, Frans Coenen, and Wenjin Lu. "Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles." *Machine vision and applications* 24, no. 7 (2013): 1405-1420.

[53] Zheng, Bichen, Sang Won Yoon, and Sarah S. Lam. "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms." *Expert Systems with Applications* 41, no. 4 (2014): 1476-1482.

Machines and Algorithms
http://www.knovell.org/mna

Research Article

Learning Scholarly Network of Influential Authors Impact of author's position on H-Index

Masood Ahmed[1, *] and Khalid Iqbal[1]

[1]COMSATS University Islamabad (Attock Campus), Attock, 43600, Pakistan
[*]Corresponding Author: Masood Ahmed. Email: me.masood.ahmed@hotmail.com

Received: 7 December 2022; Revised: 30 December 2022; Accepted: 28 February 2023; Published: 6 March 2023
AID: 002-01-000020

Abstract: There have previously been several algorithms developed for academic networks to determine the author's productivity and influence. The primary goal of these algorithms is to compute bibliometric characteristics like as publication and citation counts. There are some that are similar to the h-index, I-index, and G-Index. However, all of these are primarily concerned with citation and publication counts, and they have certain drawbacks as well. All of these algorithms are primarily used to determine the productivity and influence of authors. They most important factor which these algorithms lack is identify the contribution of an author in a research paper. Our purpose is to create author's network using the available dataset (DBLP dataset) and then build position-based algorithm which take care of author position in research paper to find out the author actual productivity in related field. Base line of this algorithm will be h-index. To calculate the points author, gain due to his/her position in paper we will consider only that paper of an author which fall inside his h-index range.

Keywords: scholarly networks; H-index; author's position; g-index; social network metrics; DBLP; R & AR indices;

1. Introduction

We have different kind of networks in computer world; some of these include social networks, scholarly networks and computer network etc. For each network there are few basic requirement-like nodes, connectivity media etc. Research articles are regarded nodes in a scholarly network of writers. Each study article contains references to other studies that are related to the one under consideration. When a paper is quoted in another publication, its relevance score rises [1], affecting the author's worth as well as the journal or conference that published the work [2]. Different mathematical principles are used to examine these writers' work, both qualitatively and quantitatively [3].

When discussing how to assess the impact of a research project, we have two key components to consider: number and quality. Quantity refers to the number of research papers published by an author, whereas quality refers to the influence of a research article in its subject. Quantity is easier to quantify than quality since quality involves many other factors [4].

https://doi.org/10.66108/mna.v2i1.35

We already have few rankings and indexing algorithms to analyze the behavior of a scholarly network. The main purpose of these algorithms is to analyze an author work and find how important their work is and what a researcher has achieved in terms of impact of their work in research domain or productivity. Comparison is made among the researchers working in same field. This comparison related algorithm is being used by different organizations to allocate scholarships and funds to academic institutions on the basis of contribution to their respective research field and their performance in this field. In following we will discuss few of the ranking and indexing algorithms already developed to assess the research performance of an author.

Hirsch's H-index [5] is regarded as the first academic networks method that is used to measure the relevance of a scientist's work. The H-index is thought to be the best technique for determining the relevance of a scientist's research output in a certain field. Hirsch works shows that it is possible to not only identify the significance of a research work but it can also tell magnitude of discoveries and efforts made by authors.

New research work is being done continuously and as time pass by all new work is added in the scholarly network world. This new research work refers to research papers of other scientist who have worked or currently working in the same field or in same area of problem c current author is working on. Logically when the work of a research scientist is referenced by other researchers in their work then the significance of referenced work should be amplified. But when we are using the Hirsh's [5] method Specifically, the H-Index does not take into account new references made to the work in other researcher's papers after a publication has already been used to quantify the h-index, implying that the h-index does not rise.

To address the h-index issue, Egghe [6] created a more efficient approach that takes into consideration freshly added references to a publication that is already in the h-index. The approach developed by Egghe [6] is known as g-index. In this technique we have an edge “g-index” over h-index in a way to show maximum useful links of a research author in more visible and clearer way.

Once g-index is done we have another issue which is consider the evaluation of work. Work done by an old researcher and a fresh researcher cannot be combined. To address this issue, the "h-index" was devised to be viewed as h-index per year value. That is, the h-index should be divided by the number of years spent by scholars in their field of study. This is known as the m-index. [7].

Then we have the R-index [8]. The purpose of this index is same as what we have in index but this index is considered to be more appropriate than “g-index”. Jon M. Kleinberg [9] proposed the Hypertext Induced Topic Selection (HITS) technique. It is consider being an ancestor of “PageRank” [10,11] algorithm. HITS use the link structure of a hypertext document as its input data to get the information related to the text of the paper. Same way the PageRank [10,11] algorithm ranks the page by getting the information from incoming links of document. To expedite the review and typesetting process, authors must follow the Microsoft Word template provided for preparing their manuscripts. This template must be strictly adhered to when formatting the manuscript for submission.

2. Literature Review

For quantifying an author's effect and productivity, a number of ranking and indexing algorithms have been created, some of which are detailed below.

2.1. H-index

At first, we have certain bibliometrics such as citation counts and paper counts. These indicators are routinely used to assess an author's success. There were several serious issues with these measures, such as the fact that the number of research publications represents the amount of effort and has nothing to do with the significance of the research activity. The amount and quality of work cannot be represented by a single numerical figure. This has a huge impact on how researchers work since it forces them to produce fewer papers in order to avoid having a low citation count. Hirsch's H-index [5] is regarded as the first academic networks method that is used to measure the relevance of research effort. The H-index is thought to be the best method for judging the relevance of a scientist's research output in a certain field. Hirsch's study

demonstrates that it is feasible to determine not only the value of a research endeavor, but also the scale of discoveries and efforts made by writers.

Hirsch [5] proposed organizing the papers in descending order based on their citation count. Hirsch [5] assigned each document a serial number R beginning with 1. The researcher's h-index is now the sequence number at which R must be larger or equal to citation counts, but R+1 must be less than citation counts.

2.2. The g-index

When utilizing the Hirsh's [5], it is crucial to remember that the h-index does not rise if new references are added to a publication after it has been included in the quantified h index in other articles. Consequently, Egghe [6] developed a more useful method to account for the recently obtained references. The g index is the name of this novel method. The g-index is superior than the h-index in that it more clearly, conspicuously, and prominently reflects the maximal helpful output of a scientist.

The publications are sorted by citation count in decreasing order in which they have been recognized in other publications. The square rank of each publication is then computed in conjunction with the cumulative sum of citations to determine the g-index. The rank at which (rank)2 is less than the cumulative total of citations in a way that makes (rank+1)2 bigger than the cumulative sum is the g-index value.

Egghe's g-index is a highly helpful variant of the pioneer index, which is the h-index, even though it is not particularly popular or has attracted academic attention for scientific output. It inherits the h-index's benefits. It is superior to the h-index in that it presents a scientist's maximal helpful output in a clearer, more conspicuous, and more apparent manner. The H-index is a straightforward and basic calculation [5].

2.3. M-Index

The H-Index is not a good tool to evaluate young scientists' research projects. Using m-index is the workaround for this h-index constraint [7]. M-quotient is another name for the m-index. The value of the h-index is divided by the researcher's years to get the m-index [5, 7].

While the m-quotient is a useful metric for evaluating the work of novice researchers based on the number of years spent conducting research, one limitation of the m-index is that a small fluctuation in the h-index value will result in a significant shift in the m-quotient value [7]. Furthermore, the author of [5] stated that the h-index should not be determined only on a scientist's initial publication, especially if the author contributed very little to the work.

2.4. R and AR indices:

Even if a publication receives twice or three times as many citations after it is listed in the h index, these citations have no bearing on the h-index. Thus, authors with the same h-index but differing citation counts cannot be distinguished by the h-index. Let's take an example where writers A and B both have two published papers and a two h-index. Paper PA2 only obtained two citations, but Author A's piece PA1 received twenty. Similarly, author B received two citations for manuscript PB2 and five for article PB1. Both writers have the same rank and influence according to the h-index, yet author A has a bigger impact than author B. To address this issue, JIN BiHui et al. develop the R-index [7]. The following formula for the R index is suggested by the authors.

$$R = \sqrt{\sum_{j=1}^{h} cit_j}$$

R-index may be used to identify various scientists that appear to be the same when looking at the h index measure but are actually not. The AR-index is the option to use when it's necessary to give newer work more relevance and older work less significance. It's computed by:

$$AR = \sqrt{\sum_{j=1}^{h} \frac{cit_j}{a_j}}$$

The formula for calculating age is (k-0.5), where k is the number of years. When it is necessary to give recent articles a higher priority than older studies, this index is utilized.

2.5. HITS:

Jon M. Kleinberg [9] introduces the concept of Hypertext Induced Topic Selection (HITS). The PageRank algorithm may be traced back to its predecessor. It exploits the link structure of hypertext documents. The underlying assumption is that the link structure of a document can provide insights into its content.

The HITS algorithm, often referred to as the Hubs and Authorities algorithm, utilizes the concept of a hub. A hub is a page that lacks substantial informative material but is rich in links to other pages. It operates in a manner similar to a directory. The pages represented by the hub are referred to as authority pages due to their provision of authoritative information on a certain issue.

The process starts by identifying a collection of sites referred to as the root set, which are pertinent to a given query. Subsequently, the hub and authority values are computed for this set of pages. In order to measure the extent of authority, the collective magnitude of the pages that reference this specific authority is aggregated. When determining the size of the hub, the authorities' magnitudes are merged.

It is crucial to acknowledge that this process relies on the ideas of in-degree and out-degree, which can lead to problems like mutual reinforcing and subject drift [9]. Both of these challenges may be resolved by assigning weights to the algorithm.

2.6. The Page Rank

Larry Page [10] and Sergey Brin [11] introduced the concept of PageRank as a means to measure the relative significance of websites on the internet. The algorithm takes into account the number of links sent towards a page, however not all inbound connections are given equal weight. The incoming connections have a role in defining the quantitative importance depending on their relative worth. This approach operates by first assigning equal weight to all pages that have outbound links to a page A, whose rank is being sought. When a page has links to both page A and other external websites, these linkages are taken into account and the initial weight assigned to page A is evenly distributed across all of these links. Only a small portion of the weight of this page is currently utilized in determining the weight of the page being evaluated.

Incoming links that possess a high rank have a greater impact or influence. Web pages inside a web environment are equipped with forward connections known as outgoing links and backward connections referred to as incoming links. Estimating the forward links of an internet page is rather simple, however locating the back-links is exceedingly challenging. The stability and immutability of links on a webpage are not obligatory. If there exists a set of pages Q(k), where each page y has a hyperlink to page k, then the formula for PageRank [12] proposed by authors is as follows:

$$PR(k) = (1 - d) + d \sum_{y=Q(k)} \frac{PR(y)}{N(y)}$$

The symbol "d" in the above equation represents the damping factor. One drawback is that at starting, all outgoing and incoming connections are given identical weight, resulting in the rank being distributed evenly among them. However, in actuality, not all links hold the same level of relevance since various linkages have varying degrees of significance. Another limitation is that it is designed for a generic internet user, but not all users fit this profile. For instance, it may not consistently yield satisfactory outcomes for a researcher. PageRank assigns greater rankings to older pages, which might be considered a disadvantage of the algorithm.

2.7. Weighted PageRank:

W. Xing et al [12] describe a weighted version of the PageRank algorithm that adds weights to pages that use both forward and backward links rather than treating all web pages equally at first. The number of forward and backward linkages provides a decent concept for assigning weights. Web sites that are more popular are given a higher weight than web pages that are less popular [13].

Both forward and backward links help figure out how popular a web page is, and a page's rank is based on how popular it is. One problem with weighted PageRank is that, like the original PageRank, it only looks at forward and backward links. It doesn't look at other things, like how semantically sound a web page is. This is how the weighted PageRank [12] is shown mathematically:

$$PR(u) = (1 + D) + d \sum_{v \in B(u)} PR(v)\, W^{in}_{(v,u)} W^{out}_{(v,u)}$$

The above equation is based on two important metrics: and. These represent the back-links to u and the forward-links to and from u, respectively. The equation then figures out the score for the page u. Here's how to figure out both and:

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p \in R(v)} Ip}$$

$$W^{out}_{(v,u)} = \frac{O_u}{\sum_{p \in R(v)} Op}$$

In the last expression, Iu and Ou show how many backlinks and forward links there are for u, and Ip and Op show how many backlinks and forward links there are for p. The author of [13] used weighted page rank to figure out how well-known experts were. The following calculation was used to figure out weighted PageRank:

$$PR_{w(pi)} = (1 - d) * \frac{w(pi)}{\sum_{k=1}^{N} w(pk)} + d \sum_{pj \in M(pi)} \frac{PR\_w(pj)}{L(pj)}$$

3. PROPOSED METHODOLOGY

This study is divided into two sections. In the first, we will use an existing index, h-index, to identify the top-ranked author in our scholarly dataset, DBLP. Once these findings are obtained, we will apply our newly developed method to determine the author's h-index rank based on his/her position in various papers.

As we all know by the standard of research paper format, an author is positioned in the paper based on his/her contribution in the research work, person who contributed most in the research work is places in first position and person who contributed least in the research work is places at the last.

The second part of this research is mainly focused on this criterion, the algorithm which we are going to build will take in consideration the position of an author. We will apply three different point gain methods to find out the author's total point gain on the basis of author position in the research papers under the consideration and then we will take the average of it to find out the actual author point gain in research papers.

Mathematically it is given below:

$$PA = \frac{\sum_{i=1}^{3} PA_i}{3} \quad (1)$$

In the above equation, (1) 1-3 represent the total number of point gain methods we will use to determine the influential authors' points in research papers, represent an author's position point gain calculated using one measure, and PA is the sum of all point gain calculated using the proposed point gain measure.

The idea of point gain rank will be applied on results of these ranking and indexing algorithms to determine that how much percentage of authors are below from the current author. In order to find out the research consistency of an author during his academic career, standard deviation will be applied.

Below are the three proposed point gain algorithms we will use the find out the author points in his/her research paper.

3.1. Authors points gain based on total authors in paper

In first author's point gain method we will assign the point to each author based on his position in total number of authors. In this system, each author receives his or her fair portion of the overall number of citations in a publication. Following is the proposed algorithm to calculate point gain of an author:

i. Select an author from the list
ii. Calculate the h-index of each article this scholar generated. Select the paper which lie within the H-index of author
iii. Iterate through each paper of author we selected in last step. Calculate the points gain by the author based on following two values
    a. Number of citations of current paper
    b. Author position in the paper
    c. Author's point will be calculated using the following formula.
        i. $P_{(g)} = C_{a*} \frac{100-(10*\sum_1^3 Pa)}{100}$
        ii. Here $P_{(g)}$ represents the point gain of an author in current paper, C_a represents the citation count of a paper, $\sum_i^{n+1} A_i$ represent the total author in a paper and P_a represent the position of author under consideration in current paper
iv. Sum all the points gain in last step.
v. Calculate the Authors total point using the following formula.

$$A_{(pg)} = \frac{\sum_i^n P(g)}{H_i}$$

Here $A_{(pg)}$ represents the author total point gain in all paper lies inside h-index, $\sum_i^n P_{(g)}$ represents the sum of point gain in all paper.

Example: Let's suppose we have a paper with citation count of 100, and the top author will receive 100 points, the second will receive 80 points, the third will receive 60 points, the fourth will receive 40 points, and the last will receive 20 points.

3.2. Author points gain based on threshold assign to position

In second author's point gain method we will assign the point to each author based on his position in total number of authors. In this methodology each author gets his/her due share with respect to total number of citations a paper has. Following is the proposed algorithm to calculate point gain of an author:

i. Select an author from the list
ii. Calculate the h-index of each article this scholar generated. Select the paper which lie within the H-index of author
iii. Iterate through each paper of author we selected in last step. Calculate the points gain by the author based on following two values
    a. Number of citations of current paper

b. Author position in the paper
c. Threshold value assign on position change
i.e., in our case we assign 5 and after position # 19 all author will get equal share
d. Author's point will be calculated using the following formula.
e.

i. $$P_{(g)} = C_{a*} \frac{100-(5*\sum_1^{19} Pa)}{100}$$

ii. Here $P_{(g)}$ represents the point gain of an author in current paper, C_a represents the citation count of a paper, $\sum_1^{19} Pa$ represent the position of author under consideration in current paper

iv. Sum all the points gain in last step.
v. Calculate the Authors total point using the following formula.

$$A_{(pg)} = \frac{\sum_i^n P(g)}{H}$$

Here $A_{(pg)}$ represents the author total point gain in all paper lies inside h-index, $\sum_i^n P_{(g)}$ represents the sum of point gain in all paper

Example: As discussed in last example we have a paper with citation count of 100, and there are 5 authors involve in this paper then first author will get 100 points, second will get 95 points, third will get 90 points, fourth will get 85 points and last one will get 80 points. Incase total authors are more than 19 then all author from 19 onwards will get only 5 points.

3.3. Author points gain based on top five positions

In third author's point gain method we will assign the point to each author based on his position in total number of authors. In this technique points are given only to top five authors after that author receive no points. Following is the proposed algorithm to calculate point gain of an author:

i. Select an author from the list
ii. Calculate the h-index of each article this scholar generated. Select the paper which lie within the H-index of author
iii. Iterate through each paper of author we selected in last step. Calculate the points gain by the author based on following two values
  i. Number of citations of current paper
  ii. Author position in the paper
  iii. Author's point will be calculated using the following formula.

i. $$P_{(g)} = C_{a*} \frac{\sum_i^{n+1} A_i}{(\sum_i^{n+1} A_i - Pa)}$$

ii. Here $P_{(g)}$ represents the point gain of an author in current paper, represents the citation count of a paper, represent the total author in a paper and represent the position of author under consideration in current paper

iv. Sum all the points gain in last step.
v. Calculate the Authors total point using the following formula.

$$A_{(pg)} = \frac{\sum_i^n P(g)}{H_i}$$

Here $A_{(pg)}$ represents the author total point gain in all paper lies inside h-index, $\sum_i^n P(g)$ represents the sum of point gain in all paper

Example: Considering the same example we discussed in last two sections for a paper with citation count of 100, and there are more than 3 authors involve in this paper then first author will get 100 points , second will get 90 point , third will get 80 points, author listed after the 5th position will get no points from that paper.

4. EXPERIMENT AND ASSESSMENT

In this part we are going to discuss the setup we used for our data collection and for experiments on the data.

4.1. Data Collection

Data for the experiment was collected from Aminer portal, following data is gather from this portal. Below is the detail of data collected from Aminer portal.

Table 1: Complete data set

Parameter	Representation	No. of records
Paper	P	2,092,356
Citations	C	8,024,869
Author	A	1,712,433
Author Collaboration	AC	4,258,615

Once this data is collected, we created a collaboration network in our database. H-Index was calculated on whole dataset. After the calculation of H-index we created another network in which we import the top 100 author on H-index to perform experiment on less data to make it in presentable form below is the detail of dataset we use for all proposed algorithms.

Table 2: Top 100 Authors data set

Parameter	Representation	No. of records
Paper	P	18,767
Citations	C	562,391
Author	A	100
Author Collaboration	AC	63, 313

4.2. Experimental setup

By using the Microsoft visual studio 2013 data was imported in Microsoft SQL server 2008 database. Once the data is imported, we executed our proposed algorithm using the SQL server query editor. For implementation the Dell system was used having processor of 2.40 GHz core i5 with 8GB RAM. We executed all three proposed algorithms and then store their result in excel sheet to generate graphs.

5. EXPERIMENTAL RESULTS

After applying all three proposed algorithms on the dataset and then calculating their average it shows that author position plays a very important role when calculating the author ranking. Results shows that a high ranked author in H-index list does not mean his contribution are also more than the other authors. A low ranked author over takes the high ranked just because of his position i.e., contribution in some influential papers.

In first step we calculated the H-index so that we know the list of influential authors we have on the basis of H-Index. Once we have calculated the list of influential authors, we created a separate network to top 100 authors.

Below table shows the top 25 influential authors from that list calculated on the bases of H-Index using the DBLP data set.

Table 3: Top 25 Authors using H-index

Author	H-Index
Hector GarciaMolina	60
Scott Shenker	56
Jiawei Han	53
David E. Culler	51
Anil K. Jain	50
Chris Faloutsos	50
Jeffrey D. Ullman	49
Ian Foster	46
P S Yu	46
Christos H. Papadimitriou	45
D Estrin	45
Hari Balakrishnan	45
Jennifer Widom	45
Jon M. Kleinberg	45
W Bruce Croft	44
Ben Shneiderman	43
Rakesh Agrawal	43
T. Anderson	43
M. Naor	42
Michael Stonebraker	42
Mihir Bellare	42
Moshe Y. Vardi	42
Pat Hanrahan	42
Tom Henzinger	42
David A. Patterson	41

Once we have the list of influential authors using the h-index our next step was to execute our proposed algorithms on top 100 influential authors to find out the impact of author positioning on his/her ranking. Following are the tables which shows the result we obtain by applying the proposed algorithm one by one

5.1. Author points gain based on total authors in paper

Table 4: Top 25 author using total author algorithms

Author	H-Index
Rakesh Agrawal	305.06
Leslie Lamport	180.55
E. M. Clarke	176.60
A Shamir	165.05

Anil K. Jain	160.87
Jiawei Han	154.01
Ian Foster	143.43
James N. Gray	136.80
Oded Goldreich	130.98
Jeffrey D. llman	129.92
Mihir Bellare	124.22
Philip A. Bernstein	122.07
Rajeev Alur	119.73
R. E. Schapire	119.53
Jon M.Kleinberg	119.29
Robert Morris	118.98
Dan Boneh	116.03
R Motwani	114.26
Serge Abiteboul	109.23
David R. Karger	109.13
Ronald Fagin	107.70
Andrew R.	107.28
S Osher	106.20
P Raghavan	102.75
I. Stoica	101.78

This list shows that the author which were part of top 25 influential author list using h-index are no more there which means that author position does have impact on the ranking of author and also it helps in calculating how much contribution an author made so far regardless of h-index. To verify our proposed research methodology, we apply our two other proposed algorithm to see how much impact they make in calculation of points gain. Table 5 & 6 the result we calculated using the “Author points gain based on threshold assign to position” and “Author points gain based on top five positions” respectively.

Table 5: Top 25 author using threshold algorithm

Author	H-Index
Rakesh Agrawal	316.24
Jeffrey D. Ullman	223.87
A Shamir	211.79
Leslie Lamport	196.03
Robert Morris	194.85
Jitendra Malik	183.03
Anil K. Jain	182.51
E. M. Clarke	180.65
Jiawei Han	178.34
Hari Balakrishnan	177.64
R Motwani	167.94
David E. Culler	163.89
Michael I. Jordan	159.21
R. E. Schapire	158.42

P Raghavan	156.83
M.Frans Kaashoe	155.83
Ian Foster	155.33
David R. Karger	148.85
S Osher	143.85
James N. Gray	143.57
Andrew Zisserman	142.35
Jon M. Kleinberg	142.28
Philip A. Bernstei	140.43
Oded Goldreich	139.96
I. Stoica	138.65

Table 6: Top 25 author using top five author algorithms

Author	H-Index
Rakesh Agrawal	295.08
Jeffrey D. Ullman	192.46
A Shamir	191.44
Leslie Lamport	181.89
E. M. Clarke	167.62
Anil K. Jain	167.02
Jiawei Han	162.08
Robert Morris	159.43
Jitendra Malik	157.89
Hari Balakrishnan	143.97
R Motwani	143.65
R. E. Schapire	141.52
Ian Foster	141.28
David E. Culler	135.51
P Raghavan	133.47
Michael I. Jordan	132.85
James N. Gray	129.73
Oded Goldreich	129.16
Philip A. Bernstei	128.80
Jon M. Kleinberg	127.97
S Osher	126.87
David R. Karger	124.94
Andrew Zisserma	122.21
M. Frans Kaashoek	121.88
Andrew R. McCallum	121.76

As we can see there is not a lot of difference in second and third algorithm but result in first algorithm are a bit different from these twos, to overcome this issue we decided to take the mean of results we achieved through these algorithm

Table 7: Top most author comparison

H-Index Top Most Author		**Position base Top Author**	
Hector Garcia-Molina		Rakesh Agrawal	
Position	Paper count	Position	Paper count
1	77	1	117
2	211	2	25
3	70	3	20
4	22	4	5
5	8	5	1
6	5	6	0
7	1	7	0
8	2	8	0
9	2	9	0
Total paper	398	**Total paper**	168

In Table 8 we listed all 25 authors and with their h1-index, points gain through three proposed algorithm and their final contribution points.

Table 8: Top 25 author complete table

Author	**HI**	**Proposed Methods A**	**Proposed Methods B**	**Proposed Methods C**	**Proposed Methods Points**
Rakesh Agrawal	43	305.06	316.24	295.08	305.46
A Shamir	34	165.05	211.79	191.44	189.43
Leslie Lamport	33	180.55	196.03	181.89	186.16
Jeffrey D. Ullman	49	129.92	223.87	192.46	182.08
E. M. Clarke	40	176.60	180.65	167.62	174.96
Anil K. Jain	50	160.87	182.51	167.02	170.13
Jiawei Han	53	154.01	178.34	162.08	164.81
Robert Morris	36	118.98	194.85	159.43	157.75
Jitendra Malik	35	99.21	183.03	157.89	146.71
Ian Foster	46	143.43	155.33	141.28	146.68
R Motwani	40	114.26	167.94	143.65	141.95
R. E. Schapire	35	119.53	158.42	141.52	139.82
Hari Balakrishnan	45	93.81	177.64	143.97	138.47
James N. Gray	33	136.80	143.57	129.73	136.70
Oded Goldreich	40	130.98	139.96	129.16	133.37
David E. Culler	51	100.46	163.89	135.51	133.29
P Raghavan	38	102.75	156.83	133.47	131.02
Philip A. Bernstein	34	122.07	140.43	128.80	130.43
Jon M. Kleinberg	45	119.29	142.28	127.97	129.85
David R. Karger	36	109.13	148.85	124.94	127.64
S Osher	35	106.20	143.85	126.87	125.64
Michael I. Jordan	35	83.53	159.21	132.85	125.20
M. Frans Kaashoek	37	88.31	155.83	121.88	122.01

6. CONCLUSION AND FUTURE WORK

A detailed analysis of effect of author position in research paper by selecting the research paper based on their h-index is covered in this paper. We have also compared the top most author of h-index and our position base ranking it clearly shows that author who top the position base ranking has a smaller number of papers as compare to h-index top author and also, he was first position author is his 2/3 of papers whereas h-index top rank author was first position author in his 1/5 of research papers see table 7. These algorithms can evolve over time, which means they can search a bigger area and change the goal threshold values on the fly. In this paper we only took the H-index as our base point to find out most effective papers in future we can use other scholarly indices to gather the affective paper and also, we can explore more on authors collaboration with each other using the author network we created.

References

[1] Mingers, John. "Measuring the research contribution of management academics using the Hirsch-index." *Journal of the Operational Research Society* 60, no. 9 (2009): 1143-1153.

[2] Yan, Erjia, and Ying Ding. "Discovering author impact: A PageRank perspective." *Information processing & management* 47, no. 1 (2011): 125-134.

[3] Katsaros, Dimitrios, Leonidas Akritidis, and Panayiotis Bozanis. "The f index: Quantifying the impact of coterminal citations on scientists' ranking." *Journal of the American Society for Information Science and Technology* 60, no. 5 (2009): 1051-1056.

[4] Kleinberg, Jon M. "Authoritative sources in a hyperlinked environment." *Journal of the ACM (JACM)* 46, no. 5 (1999): 604-632.

[5] Hirsch, Jorge E. "An index to quantify an individual's scientific research output." *Proceedings of the National academy of Sciences* 102, no. 46 (2005): 16569-16572.

[6] Egghe, Leo. "Theory and practise of the g-index." *Scientometrics* 69, no. 1 (2006): 131-152.

[7] Harzing, Anne-Wil, and Ron van der Wal. "A Google Scholar H-Index for journals: A better metric to measure journal impact in economics & business." In *Proceedings of the Academy of Management Annual Meeting*. Wiley, 2008.

[8] Jin, Bihui, LiMing Liang, Ronald Rousseau, and Leo Egghe. "The R-and AR-indices: Complementing the h-index." *Chinese science bulletin* 52, no. 6 (2007): 855-863.

[9] Kleinberg, Jon M. "Authoritative sources in a hyperlinked environment." *Journal of the ACM (JACM)* 46, no. 5 (1999): 604-632.

[10] Page, Larry, Sergey Brin, Rajeev Motwani, and Terry Winograd. *PageRank: Bringing order to the web*. Vol. 72. Stanford Digital Libraries Working Paper, 1997.

[11] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine." *Computer networks and ISDN systems* 30, no. 1-7 (1998): 107-117.

[12] Xing, Wenpu, and Ali Ghorbani. "Weighted pagerank algorithm." In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pp. 305-314. IEEE, 2004.

[13] Ding, Ying. "Applying weighted PageRank to author citation networks." *Journal of the American Society for Information Science and Technology* 62, no. 2 (2011): 236-245.