**Palestinian Journal**
**of Technology & Applied Sciences (PJTAS)**

**Palestinian Journal**
**of Technology & Applied Sciences (PJTAS)**

## Vision

Achieving leadership, excellence and innovation in the field of open learning, community service, and scientific research, in addition to reinforcing the University leading role in establishing a Palestinian society built on knowledge and science.

## Mission

To prepare qualified graduates equipped with competencies that enable them to address the needs of their community, and compete in both local and regional labor markets. Furthermore, The University seeks to promote students' innovative contributions in scientific research and human and technical capacity-building, through providing them with educational and training programs in accordance with the best practices of open and blended learning approach, as well as through fostering an educational environment that promotes scientific research in accordance with the latest standards of quality and excellence. The University strives to implement its mission within a framework of knowledge exchange and cooperation with the community institutions and experts.

## Core Values

*To achieve the University's vision, mission and goals, the University strives to practice and promote the following core values:*

- Leadership and excellence.
- Patriotism and nationalism.
- Democracy in education and equal opportunities.
- Academic and intellectual freedom.
- Commitment to regulations and bylaws.
- Partnership with the community
- Participative management.
- Enforcing the pioneer role of women.
- Integrity and Transparency.
- Competitiveness.

## The Journal

The Palestinian Journal of Technology and Applied Sciences is an annual scientific refereed journal, issued by the Deanship of Graduate Studies and Scientific Research. The first issue of the Journal was published in January 2018 after obtaining an International Standard Serial Number (E- ISSN: 2521-411X), (P– ISSN: 2520-7431).

The journal publishes original research papers and studies conducted by researchers and faculty staff at QOU and by their counterparts at local and overseas universities, in accordance with their academic specializations. The Journal also publishes reviews, scientific reports and translated research papers, provided that these papers have not been published in any conference book or in any other journal.

**The Journal comprises the following topics:**

Information and Communication Technology, Physics, Chemistry, Biology, Mathematics, Statistics, Biotechnology, Bioinformatics, Agriculture Sciences, Geology, Ecology, Nanotechnology , Mechatronics, Internet of things , Artificial Intelligence and Big Data.

Palestinian Journal
of Technology & Applied Sciences (PJTAS)

# Publication and Documentation Guidelines

## *First: Requirements of preparing the research:*

***The research must include the following:***

1. A cover page which should include the title of the research stated in English and Arabic, including the name of researcher/researchers, his/her title, and email.
2. Two abstracts (English and Arabic) around (150-200 word). The abstract should include no more than 6 key words.
3. Graphs and diagrams should be placed within the text, serially numbered, and their titles, comments or remarks should be placed underneath.
4. Tables should be placed within the text, serially numbered and titles should be written above the tables, whereas comments or any remarks should be written underneath the tables

## *Second: Submission Guidelines:*

1. The Researcher should submit a letter addressing the Head of Editorial Board in which he/she requests his paper to be published in the Journal, specifying the specialization of his/her paper.
2. The researcher should submit his research via email to the Deanship of Scientific research (tas@qou.edu) in Microsoft Word Format, taking into Consideration that the page layout should be two columns.

   (Check the attached digital form on the website of the Journal)
3. The researcher should submit a written pledge that the paper has not been published nor submitted for publishing in any other periodical, and that it is not a chapter or a part of a published book.
4. The researcher should submit a short Curriculum Vitae (CV) in which she/he includes full name, workplace, academic rank, specific specialization and contact information (phone and mobile number, and e-mail address).
5. Complete copy of the data collection tools (questionnaire or other) if not included in the paper itself or the Annexes.
6. No indication shall be given regarding the name or the identity of the researcher in the research paper, in order to ensure the confidentiality of the arbitration process.

### *Third- Publication Guidelines:*

The editorial board of the journal stresses the importance of the full compliance with the publication guidelines, taking into note that research papers that do not meet the guidelines will not be considered, and they will be returned to the researchers for modification to comply with the publication guidelines.

1. Papers are accepted in English only, and the language used should be well constructed and sound.
2. The researcher must submit his/her research via email (tas@qou.edu )in Microsoft Word format, taking into consideration the following:
   - Font type should be Times New Roman, and the researcher should use bold font size 14 for head titles, bold font size 13 for subtitles, font size 12 for the rest of the text, and font size 11 for tables and diagrams.
   - the text should be single-spaced
   - Margins: Should be set to: 2cm top, 2.5 cm bottom, 1.5 cm left and right.
3. The paper should not exceed 25 (A4) pages or (7000) words including figures and graphics, tables, endnotes, and references, while annexes are inserted after the list of references, though annexes are not published but rather inserted only for the purpose of arbitration.
4. The research has to be characterized by originality, neutrality, and scientific value.
5. The research should not be published or submitted to be published in other journals, and the researcher has to submit a written acknowledgment that the research has never been published or sent for publication in other journals during the completion of the arbitration process. In addition, the main researcher must acknowledge that he/she had read the publication guidelines and he/she is fully abided by them.
6. The research should not be a chapter or part of an already published book.
7. Neither the research nor part of it should be published elsewhere, unless the researcher obtains a written acknowledgement from the Deanship of Scientific Research.
8. The Journal preserves the right to request the researcher to omit, delete, or rephrase any part of his/her paper to suit the publication policy. The Journal has also the right to make any changes on the form/ design of the research.
9. The research must include two research abstracts, one in Arabic and another in English of (150-200) words. The abstract must underline the objectives of the paper, statement of the problem, methodology, and the main conclusions. The researcher is also to provide no more than six keywords at the end of the abstract which enable an easy access in the database.

11. The researcher has to indicate if his research is part of a master thesis or a doctoral dissertation as he/she should clarify this in the cover page, possibly inserted in the footnote.
12. The research papers submitted to the Deanship of Scientific Research will not be returned to the researchers whether accepted or declined.
13. In case the research does not comply with the publication guidelines, the deanship will send a declining letter to the researcher.
14. Researchers must commit to pay the expenses of the arbitration process, in case of withdrawal during the final evaluation process and publication procedures.
15. The researchers will be notified of the results and final decision of the editorial board within a period ranging from three to six months starting from the date of submitting the research.

### *Four- Documentation:*

1. Footnotes should be written at the end of the paper as follows; if the reference is a book, it is cited in the following order, name of the author, title of the book or paper, name of the translator if any or reviser, place of publication, publisher, edition, year of publishing, volume, and page number. If the reference is a journal, it should be cited as follows, author, paper title, journal title, journal volume, date of publication and page number.
2. References and resources should be arranged at the end of the paper in accordance to the alphabetical order starting with the surname of author, followed by the name of the author, title of the book or paper, place of publishing, edition, year of publication, and volume. The list should not include any reference which is not mentioned in the body of the paper.
   - In case the resource is with no specified edition, the researcher writes ( N.A)
   - In case the publishing company is in not available, the researcher writes (N.P)
   - In case there is no author, the researcher writes (N.A)
   - In case the publishing date is missing , the researcher writes (N.D)
3. In case the researcher decides to use APA style for documenting resources in the text, references must be placed immediately after the quote in the following order, surname of the author, year of publication, page number.
4. Opaque terms or expressions are to be explained in endnotes. List of endnotes should be placed before the list of references and resources

   Note: for more information about using APA style for documenting please check the following link:

   ***http://journals.qou.edu/recources/pdf/apa.pdf***

### *Five: Peer Review & Publication Process:*

All research papers are forwarded to a group of experts in the field to review and assess the submitted papers according to the known scientific standards. The paper is accepted after the researcher carries out the modifications requested. Opinions expressed in the research paper solely belong to their authors not the journal. The submitted papers are subject to initial assessment by the editorial board to decide about the eligibility of the research and whether it meets the publication guidelines. The editorial board has the right to decide if the paper is ineligible without providing the researcher with any justification.

***The peer review process is implemented as follows:***

1. The editorial board reviews the eligibility of the submitted research papers and their compliance with the publication guidelines to decide their eligibility to the peer review process.
2. The eligible research papers are forwarded to two specialized Referees of a similar rank or higher than the researcher. Those Referees are chosen by the editorial board in a confidential approach, they are specialized instructors who work at universities and research centers in Palestine and abroad.
3. Each referee must submit a report indicating the eligibility of the research for publication.
4. In case the results of the two referees were different, the research is forwarded to a third referee to settle the result and consequently his decision is considered definite.
5. The researcher is notified by the result of the editorial board within a period ranging from three to six months starting from the date of submission. Prior to that, the researcher has to carry out the modifications in case there are any.

### *Six: Scientific Research Ethics:*

***The researcher must:***

1. Commit to high professional and academic standards during the whole process of conducting research papers, from submitting the research proposal, conducting the research, collecting data, analyzing and discussing the results, and to eventually publishing the paper. All must be conducted with integrity, neutralism and without distortion.

2. Acknowledge the efforts of all those who participated in conducting the research such as colleagues and students and list their names in the list of authors, as well as acknowledging the financial and morale support utilized in conducting the research.
3. Commit to state references soundly, to avoid plagiarism in the research.
4. Commit to avoid conducting research papers that harm humans or environment. The researcher must obtain in advance an approval from the University or the institutions he/she works at, or from a committee for scientific research ethics if there is any, when conducting any experiments on humans or the environment.
5. Obtain a written acknowledgement from the individual/individuals who are referred to in the research, and clarify to them the consequences of listing them in the research. The researcher has also to maintain confidentiality and commit to state the results of his/her research in the form of statistical data analysis to ensure the confidentiality of the participating individuals.

### *Seven: Intellectual Property Rights:*

1. The editorial board confirms its commitment to the intellectual property rights
2. Researchers also have to commit to the intellectual property rights.
3. The research copyrights and publication are owned by the Journal once the researcher is notified about the approval of the paper. The scientific materials published or approved for publishing in the Journal should not be republished unless a written acknowledgment is obtained by the Deanship of Scientific Research.
4. Research papers should not be published or republished unless a written acknowledgement is obtained from the Deanship of Scientific Research.
5. The researcher has the right to accredit the research to himself, and to place his name on all the copies, editions and volumes published.
6. The author has the right to request the accreditation of the published papers to himself.

**Palestinian Journal**
**of Technology & Applied Sciences (PJTAS)**
No. 8

# Contents

**Palestinian Journal of Technology and Applied Sciences No – (8). June 2025**
**https://journals.qou.edu/index.php/PJTAS**
**E-ISSN: 2521-411X/ P-ISSN: 2520-7431**

# Distribution Network Performance Enhancement Using Reconfiguration Technique based on Gravitational Search Algorithm

***Dr. Ola Subhi Badran****

Department of Electrical Engineering-Industrial Automation, Faculty of Engineering and Technology, Palestine Technical University – Kadoorie (PTUK), Tulkarm, Palestine.

**Oricd No**: 0000-0002-1602-2496
**Email**: o.badran@ptuk.edu.ps

**Received:**
7/03/2024

**Revised:**
7/03/2024

**Accepted:**
26/03/2024

***Corresponding Author**:
o.badran@ptuk.edu.ps

**Citation**: Bdran, O. S. Distribution Network Performance Enhancement Using Reconfiguration Technique based on Gravitational Search Algorithm. Palestinian Journal of Technology and Applied Sciences (PJTAS), 1(8). https://doi.org/10.33977/2106-000-008-001



- **Open Access**



**Abstract**

**Objectives**: The main goals of this work are to minimize network power loss and enhance the system's voltage profile (VF).

**Methods**: This work presents a novel methodology that simultaneously optimizes Distribution Network Reconfiguration (DNR), Distributed Generation (DG) sizing, and DG placement using the Gravitational Search Algorithm (GSA) optimization technique. The DNR approach helps reduce power loss, but its effectiveness is limited when applied alone. Similarly, optimizing DG sizing and placement can further minimize power loss, but improper integration with DNR may lead to increased power loss and voltage fluctuations. Hence, it is essential to develop an efficient optimization strategy that simultaneously determines the optimal DG size and location while achieving optimal DNR.

**Results**: For the IEEE 33-bus network, active and reactive power losses were reduced by 67.488% and 64.88%, respectively. Similarly, for the IEEE 69-bus network, the reductions in active and reactive power losses were 82.55% and 62.25%, respectively.

**Conclusions**: The findings show that adjusting the size and location of distributed generation units (DGs) while configuring the network significantly improves the voltage profile and reduces losses.

**Keywords**: Gravitational Search Algorithm, Optimization Technique, Voltage Profile, Network Reconfiguration, Power Loss.

# تحسين أداء شبكة التوزيع الكهربائية باستخدام تقنية إعادة التشكيل بالاعتماد على خوارزمية الجاذبية

***د. علا صبحي بدران****

قسم الهندسة الكهربائية–الأتمتة الصناعية، كلية الهندسة والتكنولوجيا، جامعة فلسطين التقنية– خضوري، طولكرم، فلسطين.

**الملخص**

**الأهداف:** أهداف هذا العمل الرئيسية هي تقليل خسارة الطاقة في الشبكة وتعزيز منحى الجهد الكهربائي.

**المنهجية:** يوفر هذا العمل منهجية جديدة تحقق إعادة تشكيل مثلى لشبكة التوزيع (DNR) والتحديد الامثل لحجم وموقع مولدات التوزيع باستخدام تقنية خوارزمية بحث الجاذبية .(GSA) تُستخدم تقنية إعادة تشكيل شبكة التوزيع (DNR) لتقليل فقد الطاقة. ومع ذلك، فإن تطبيق هذه الطريقة وحدها سيقلل من فقد الطاقة إلى قيمة محددة. التقنية الأخرى المستخدمة لتقليل فقد الطاقة هي تحديد حجم وموقع مولدات التوزيع. ومع ذلك، فإن استخدام تقنية إعادة تشكيل المشتركة مع تحديد حجم وموقع مولدات التوزيع بطريقة غير مثلى قد يؤدي إلى زيادة فقد الطاقة وتباين الجهد. لذلك، فمن المهم تطوير منهجية تحسين فعالة تحدد حجم وموقع مولدات التوزيع المثلى وتضمن إعادة تشكيل مثلى للشبكة في نفس الوقت.

**النتائج:** تم تقليل خسارة الطاقة الفعالة وغير الفعالة بنسبة 67.488% و64.88% على التوالي لشبكة-33 IEEE bus. بينما تم تقليل فقد الطاقة الفعالة وغير الفعالة بنسبة 82.55% و62.25% على التوالي لشبكة IEEE 69-bus.

**الخلاصة:** أثبتت النتائج أن دمج عملية التحديد المثلى لحجم وموقع مولدات التوزيع مع عملية إعادة التشكيل المثلى للشبكة بشكل متزامن يؤدي إلى تحسين ملحوظ في منحنى الجهد الكهربائي والحد الأدنى للخسائر.

**الكلمات الدالة:** خوارزمية بحث الجاذبية، تقنية التحسين، منحنى الجهد، إعادة تشكيل شبكة التوزيع، فقدان الطاقة.

## Introduction

In today's distribution networks (DN), power loss is a significant concern due to the growing demand for electricity. According to the companies of electrical distribution, it may cause increased operating costs and decrease the voltage profile of the network (Yan, Shamim, Chou, Desideri, & Li, 2017). Therefore, different methods are studied by researchers to solve the electrical distribution network problems (Ola Badran, Mekhilef, Mokhlis, & Dahalan, 2017).

Power loss in distribution networks (DN) is a critical challenge in power systems. Network reconfiguration is one of the most effective methods for reducing power loss and improving reliability indices (Nguyen & Truong, 2015). This technique involves altering the status of switches to relieve network overload and minimize power loss. In (Abdelaziz, 2017), The authors introduced a novel approach to solving the Reconfiguration problem utilizing GA. This algorithm effectively handles the non-linear constraints and complex combinations associated with the reconfiguration proces. A discrete form of PSO was used in (Sivanagaraju, Rao, & Raju, 2008) for load balancing during DNR. Similar to this, GA was used in (Eldurssi & O'Connell, 2014) to solve the DNR issue with the goals of lowering power losses, raising the load index, and raising the VF. Network reconfiguration (NR) was used in (Kashem, Ganapathy, & Jasmon, 2000) to improve the VF of the radial network and maximize loadability, both of which boosted network reliability. Additionally, NR was utilized in a two-stage algorithm in (Tyagi, Verma, & Bijwe, 2018) to lower reactive power loss (REC) and enhance loadability. In (Pegado, Ñaupari, Molina, & Castillo, 2019), The authors introduced an alternative methodology to solve the DNR problem using Binary Particle Swarm Optimization (BPSO). They proposed a novel sigmoid function to enhance result convergence and regulate the rate of change in particles. The obtained results demonstrated high efficiency and reliability in identifying the optimal solution.

Voltage profile is also an important issue in the distribution system. Therefore, DG units are incorporated into the network system (Avchat & Mhetre, 2020). Thus, it is used to limit the major central power plants at peak loads and improve the system's reliability and stability (Karunarathne, Pasupuleti, Ekanayake, & Almeida, 2021). In (Moradi & Abedini, 2012), GA and PSO algorithms were proposed to find the best DG sizing (DGS) and DG location (DGL). In (O Badran & Jallad, 2014), storage batteries were integrated to the network with renewable DGs to enhance the VF. In (Mohandas, Balamurugan, & Lakshminarasimman, 2015), the authors presented a new approach to modify the voltage stability of the network. In (Ola Badran, 2023), the author used the FA algorithm to reduce power loss and improve the system voltage. Additionally, by building renewable energy DGs, (Yang et al., 2021) managed pollution emissions, power outages, DG costs, and VF in addition to addressing the issues of DG energy consumption and environmental pollution.. An algorithm combining PSO and GA was introduced in (Ha, Nazari-Heris, Mohammadi-Ivatloo, & Seyedi, 2020) to minimize active (ACT) and reactive (REC) power losses while improving voltage management. Moreover, to measure ACT and REC power loss, increase voltage stability, and boost power system security and dependability, a differential evolution method and voltage stability index were created in (Karuppiah, 2021).

Thus, DNR, DGS, and DGL techniques were combined to improve the system performance. In (Rao, Ravindra, Satish, & Narasimham, 2012), the authors solve the DGS, DGL, and DNR simultaneously to reduce power loss and enhance VF. The Harmony Search Algorithm (HSA) was used to solve the proposed problem. The obtained results were effective. Moreover, in (Imran, Kowsalya, & Kothari, 2014) a new methodology utilizing the FWA was introduced to solve the DNR and DG location problem, aiming to enhance system stability and reduce power loss. The simulated results validated the effectiveness of the proposed technique. Furthermore, (Ola Badran, Mokhlis, Mekhilef, & Dahalan, 2018), the authors minimize network power loss, reduce DG output, and improve the voltage profile (VF) index. Various metaheuristic algorithms were utilized, and the results successfully validated the effectiveness of the proposed approach.. While in (Ola Badran & Jallad, 2023a), the authors integrated a shunt capacitor to the network to reduce losses in both ACT and REC power, as well as to improve the VF by applying a multi-objective decision-making technique.

Unlike previous studies, the main contribution of this paper is the simultaneous optimization of Distribution Network Reconfiguration (DNR), Distributed Generation Sizing (DGS), and Distributed Generation Location (DGL) using the Gravitational Search Algorithm (GSA).

## 2. Objective Fitness and constraints

The proposed methodology aims to achieve optimal reconfiguration while simultaneously determining the best DG sizing and location.

The fitness is defined as the active power loss ($P_{loss}$):

$$F = (P_{loss}) \quad (1)$$

The power loss is given by:

$$P_{loss} = \sum_{N=1}^{M} (R_N \times |I_N|^2) \quad (2)$$

where N is the branch number, RN is the resistance in branch N, and IN is the branch current.
The presented method must fulfill these constraints:

**1. The DG output Capacity ($P_{DG}$):**

$$P_i^{min} \leq P_{DG,i} \leq P_i^{max} \quad (3)$$

where the allowable upper and lower bounds of the DG are denoted by $P_i^{max}$and $P_i^{min}$, respectively.

**2. Injection Power:**

$$\sum_{i=1}^{k} P_{DG,i} < (P_{Load} + P_{loss}) \quad (4)$$

where $P_{Load}$: is the power load. This constraint is meant to stop power from returning to the grid from the DG units, as that can cause problems with protection..

**3. Balance power:**

$$\sum_{i=1}^{k} P_{DG,i} + P_{Substation} = P_{Load} + P_{loss} \quad (5)$$

where $P_{Substation}$ the main substation active power. Both power load and power supply must be equal. This limitation guarantees the equilibrium principle, necessitating a balance between the supply and demand of power. In other words, the aggregate power produced by the DG units and substation needs to match the total of the power load and the power loss.

**4. Magnitude Voltage**

$$V_{min} \leq V_{bus} \leq V_{max} \quad (6)$$

where VMin and VMax are the voltage minimum (Min) and maximum (Max) values, respectively, and Vbus is the voltage bus. The range of any bus voltage must be 0.95 p.u. to 1.05 p.u. (±5% of the rated value) (Rahim et al., 2019).

**5. Configuration Form:**

The most significant restriction is that the distribution network must continue to be configured radially. (Ola Badran & Jallad, 2023c).

**6. Isolation load:**

Ensured that all nodes are energized to connect power to each node.

## 3. Proposed methodology

The optimization process involves reconfiguring the network while simultaneously determining the optimal DG sizing and placement using the Gravitational Search Algorithm (GSA). The GSA is a stochastic search technique that models agent interactions using the law of gravity to solve optimization issues. Within GSA, agents are viewed as objects whose properties are dictated by their gravitational force and masses attracting objects in the direction of larger masses, which directs the system's global motion. The following are the steps to apply the suggested GSA to DGS, DGL, and DNR:

Step 1: The parameters that make up the input data are defined, including the voltage, resistance, and reactance values of the lines as well as the bus load. The number of masses ($N_{mass}$) is the set up parameter through the GSA.

Step 2: By choosing switches at random to open in the distribution network and figuring out the size and placement of the DGs to form the masses, the first population is created. The first portion of the mass, which represents network reconfiguration, will have a length of $N_{opened}$ if the number of switches that need to be opened is $N_{opened}$. In a similar vein, $N_{DG}$, the second component of mass represents the quantity of DGs that must be added to the distribution system. In the simultaneous case, the switches to be opened and the DG sizes are configured as follows:

$$Mass_i = \left[S_1^1, S_2^2, \cdots, S_{N_{opened}}^d, D_{\mathrm{L1}}^{d+1}, D_{\mathrm{L2}}^{d+2}, \cdots, D_{LN_{DG}}^{N_d}, D_{\mathrm{S1}}^{d+1}, D_{\mathrm{S2}}^{d+2}, \cdots, D_{SN_{DG}}^{N_d}\right] \quad (7)$$

where $i = 1, 2, \cdots, N_{mass}, N_d$ is the variables or dimensions to be optimized, and $Mass_i$ is the position of $i-th$ mass in the $d-th$ dimension. $S_1^1, S_2^2$ and $S_{N_{opened}}^d$ are the opened switched in $d-th$ dimension, and $D_{\mathrm{L1}}^{d+1}, D_{\mathrm{L2}}^{d+2}$ and $D_{LN_{DG}}^{N_d}$ are the location of the DG units, and $D_{\mathrm{S1}}^{d+1}, D_{\mathrm{S2}}^{d+2}$ and $D_{SN_{DG}}^{N_d}$ are the sizes of the DG units in MW of $d-th$ dimension.

Step 3: To determine the bus's voltages and the power flow across each network line, start the first iteration by executing the power flow program. You may then calculate the power losses and the lowest voltage across all buses using these data.

$$F_i^d(iter) = \sum_{j \in Kbest, j \neq i}^{N_{mass}} rand_j F_{ij}^d(iter) \quad (17)$$

In GSA, a random number within the interval [0, 1], represented as $rand_j$, is introduced. The algorithm should gradually enhance exploitation while decreasing exploration to strike a balance between the two. By the beginning of the algorithm, every mass exerts force on every other mass, but by the conclusion, only one mass remains in contact with the others. This is accomplished by introducing the idea of $Kbest$, a function of iteration. The collection of the first $K$ masses with the biggest mass and the least power loss is denoted as $Kbest$. The initial value of $Kbest$, or $K_0$, is set at the beginning and lowers as the number of iterations increases, causing $K$best to decrease linearly over time. The next velocity of a mass is given by:

$$V_i^d(iter + 1) = rand_i \times v_i^d(iter) + a_i^d(iter) \quad (18)$$

Step 6: Update the masses' positions as indicated below:

$$Mass_i^d(iter + 1) = Mass_i^d(iter) + v_i^d(iter + 1) \quad (19)$$

Step 7: Till the maximum number of iterations is achieved, carry out the actions from step 3 again.

Step 8: When the allotted number of iterations is reached, end the procedure and produce the best result, which includes the voltage at each bus, the locations and sizes of the DGs, the switch numbers defining the new network configuration, the power losses for the process, and the corresponding fitness value.

Figure. 1 illustrates the flowchart of the proposed methodology utilizing the Gravitational Search Algorithm (GSA).

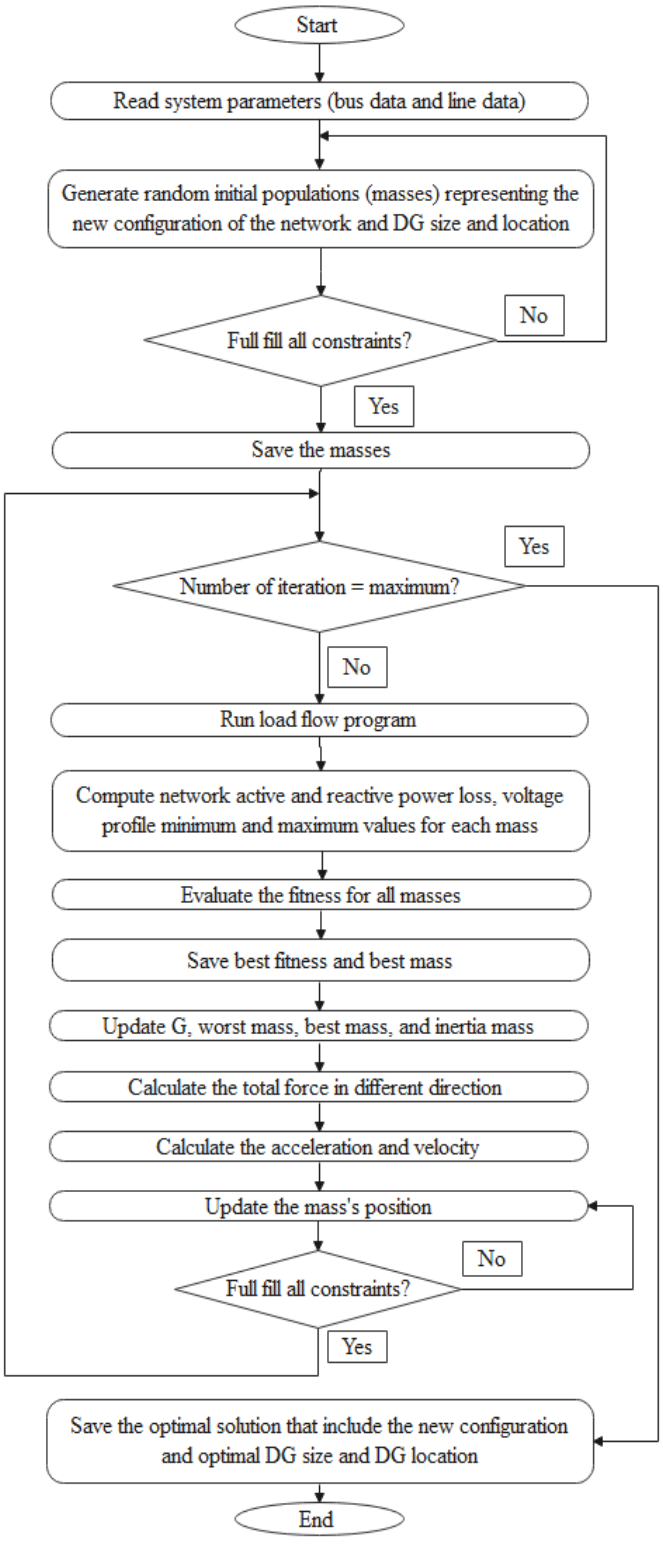



**Figure (1): Flow chart of DNR and DG_LS using GSA**

## 4. Simulation Results and Discussion

The results of the suggested approach for concurrently addressing the DG location, sizing, and system reconfiguration issues are shown in this section. MATLAB was used to implement and solve the methodology because of its robustness and speed. Every code was ran 20 times for 100 population sizes and 300 iterations during the simulation, which was executed on a laptop equipped with an Intel Core i7 processor. Two IEEE 33-bus systems (Figure 2) and an IEEE 69-bus system (Figure 3) were used to test the methodology. There were 37 switches in the IEEE 33-bus distribution

network: 32 sectionalizing switches and 5 tie switches. As shown in Figure 2, the original network had switches 33, 34, 35, 36, and 37 that were generally open and the other switches that were normally closed. The voltage of the system was 12.66 kV, and the total real load demand was 3715 kW. There was 100 MVA as the base apparent power value. At the beginning, the network experienced 202.677 kW of power losses, with 0.913 pu being the lowest bus voltage. Switches 69 through 73 were initially left open in the IEEE 69-bus distribution network, which included 73 branches, 5 tie switches, and 68 sectionalizing switches. The minimum voltage magnitude of the system is 0.9092 p.u., while its nominal voltage is 12.66 kV. The system's apparent power demand is (3,802.19 + j2, 694.6) kVA, with corresponding ACT and REC power losses of 39.16 kVAR and 224.99 kW. The complete bus and line data are given in (Ola Badran & Jallad, 2023b; Ola Badran, Jallad, Mokhlis, & Mekhilef, 2020). It was supposed that the DG in this test setup was a mini-hydro generator. Every DG had a capacity of 2 MW.

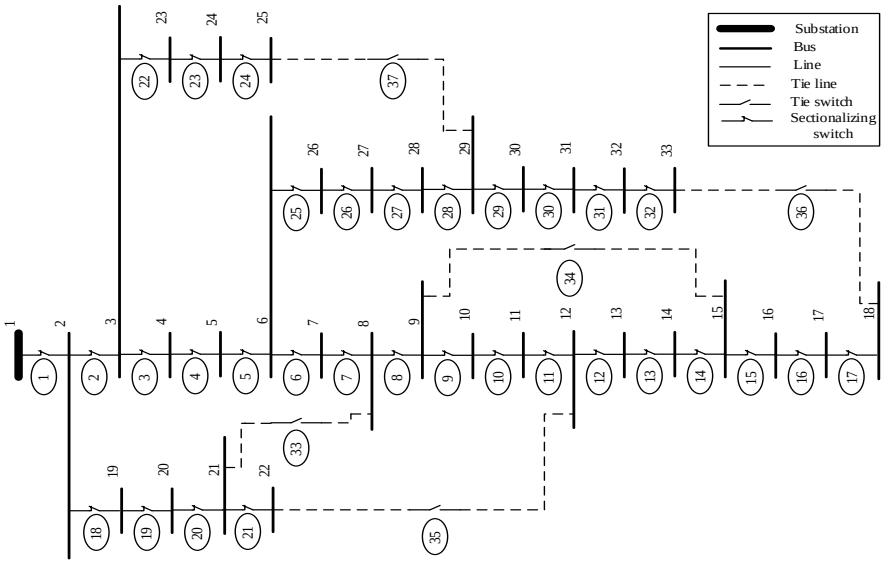



**Figure (2): IEEE 33-bus system**

### 4.1 Effect of the Optimization Technique on Power Loss

Table 1 illustrates the output results obtained by using GSA and compared to the initial form for the IEEE 33 bus system. It's seen that the optimization technique provides a better result according to the initial form. The active power loss was 65.87 kW compared to the initial case of 202.6 kW so the active power loss reduction was 67.488 %. The reactive power loss is 47.41 kVAR compared to the initial case of 135 kVAR and so the reactive power loss reduction is 53.52 %. The minimum voltage value of the voltage profile (VF) is 0.9695 p.u., an improvement over the initial case, where the minimum voltage was 0.913 p.u. Additionally, Table 2 presents the output results obtained using GSA, compared to the initial values for the IEEE 69-bus system. It's seen that the optimization technique provides a better result according to the initial form. The active power loss was 39.20 kW compared to the initial case of 224.6 kW so the active power loss reduction was 82.55 %. The reactive power loss is 38.5 kVAR compared to the initial case of 102 kVAR and so the reactive power loss reduction is 62.25 %. The minimum voltage value of the VF is 0.9807 p.u compared to the initial case where the minimum voltage is 0.9093 p.u.

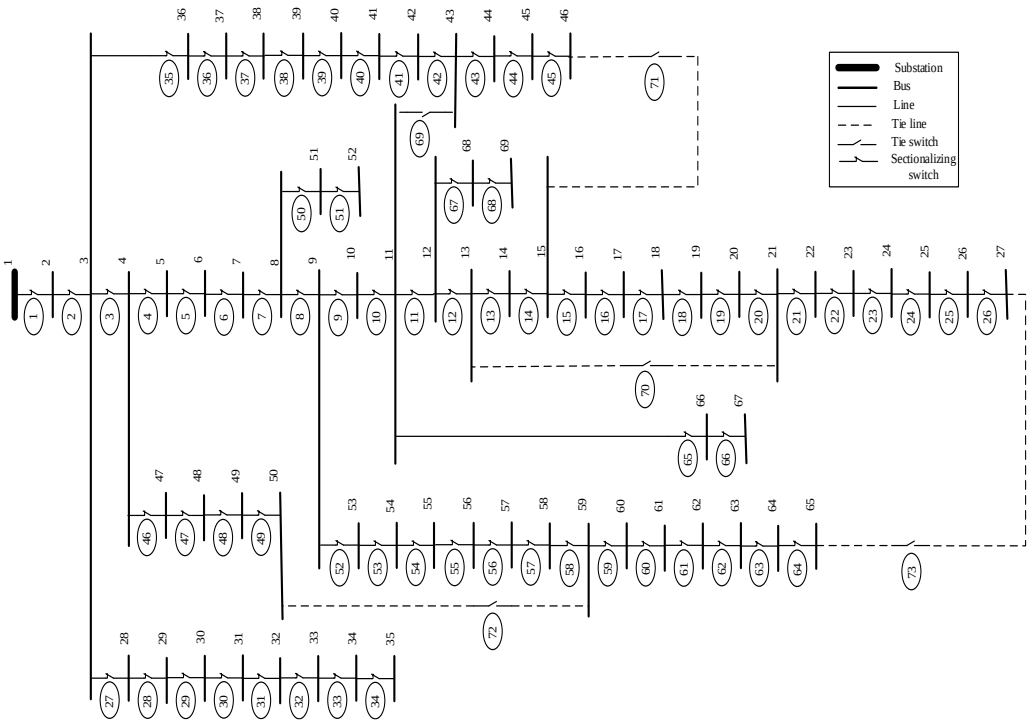



**Figure (3): IEEE 69-bus system**

**Table 1: Proposed method results for IEEE 33 bus system**

| Item | Initial Form | DNR and DG_S and DG_L Form |
|---|---|---|
| Tie switch | 33, 34, 35, 36, 37 | 32, 34, 27, 33, 9 |
| | | 32 |
| DG Location | --- | 25 |
| | | 13 |
| | | .541 |
| DG Sizing (MW) | --- | .656 |
| | | .625 |
| Fitness P_loss (kW) | 202.6 | 65.87 |
| Q_loss (kVAR) | 135 | 47.41 |
| P_loss (%) | --- | 67.488 |
| Q_loss (%) | --- | 64.88 |
| Min Voltage (p.u) | .9132 | .9695 |
| Max Voltage (p.u) | 1 | 1 |

**Table 2: Proposed method results for IEEE 69 bus system**

| Item | Initial Form | DNR and DG_S and DG_L Form |
|---|---|---|
| Tie switch | 69, 70, 71, 72, 73 | 13, 12, 10, 57, 62 |
| DG Location | --- | 22, 16, 61 |
| DG Sizing (MW) | --- | .507, .41, 1.512 |
| Fitness P_loss (kW) | 224.6 | 39.20 |

| Item | Initial Form | DNR and DG_S and DG_L Form |
|---|---|---|
| Q_loss (kVAR) | 102 | 38.5 |
| P_loss (%) | --- | 82.55 |
| Q_loss (%) | --- | 62.25 |
| Min Voltage (p.u) | .9093 | .9807 |
| Max Voltage (p.u) | 1 | 1 |

### 4.2 Effect of the Optimization Technique on Voltage Profile

The voltage profiles after applying the optimization technique are displayed in Figure 4 for the IEEE 33-bus distribution network and Figure 5 for the IEEE 69-bus distribution network. The voltage magnitudes of all buses show significant improvement compared to the initial case. After network reconfiguration, along with optimal DG location and sizing, all bus voltages are closer to unity.

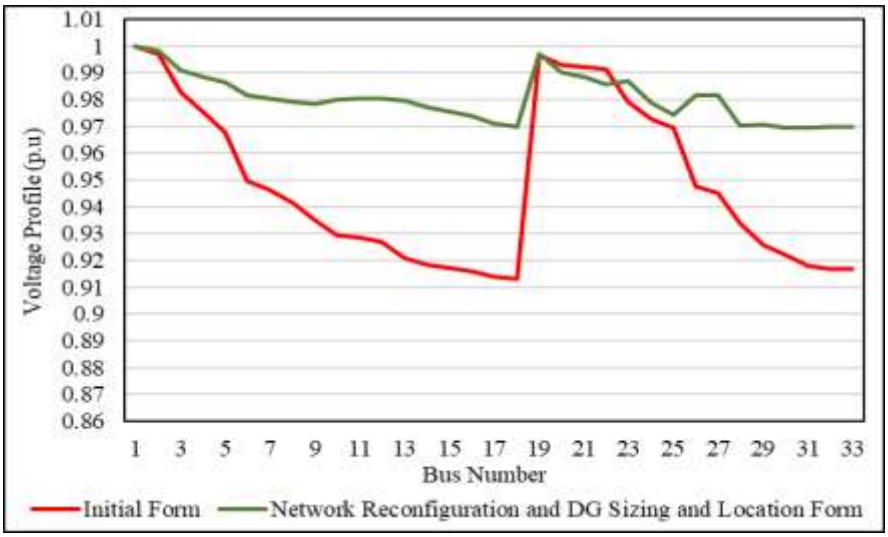



**Figure (4): IEEE 33-bus distribution network voltage profile**

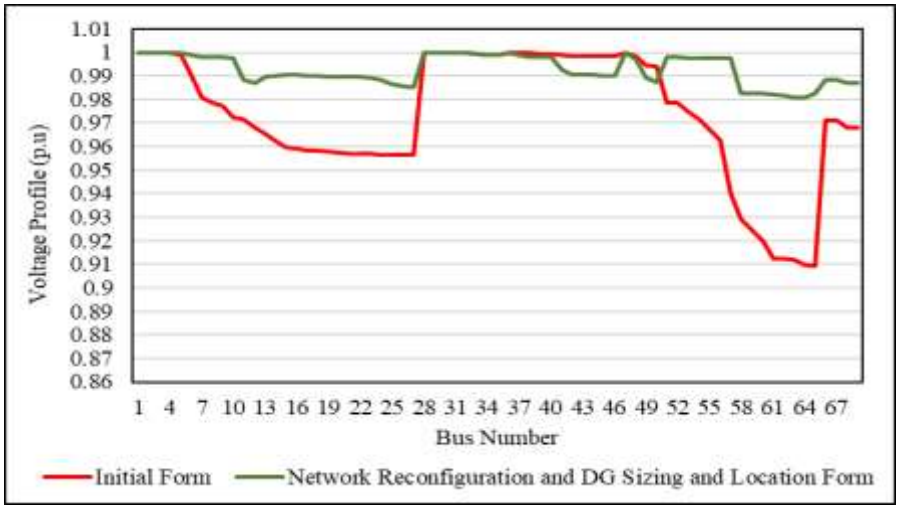



**Figure (5): IEEE 69-bus distribution network voltage profile**

### 4.3 The overall performance of GSA

Additionally, Figures 6 and 7 (corresponding to the IEEE 33-bus and IEEE 69-bus distribution networks, respectively) show results of a robustness test. The optimization method was used 20 times in this test. Every run of the GSA generated findings that were consistently similar and resulted in a locally optimal solution. The global optimal solution was shown to be the best of these local optima. In addition to DG sizing and location, the GSA showed significant and robust convergence performance, demonstrating its efficacy in identifying both local and global optimal solutions for complex problems like DNR.

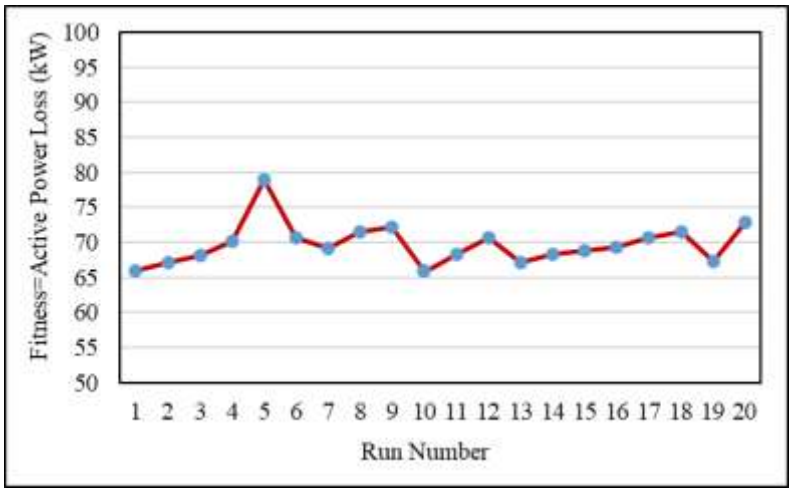



**Figure (6): GSA robustness test curve for IEEE 33 bus system**

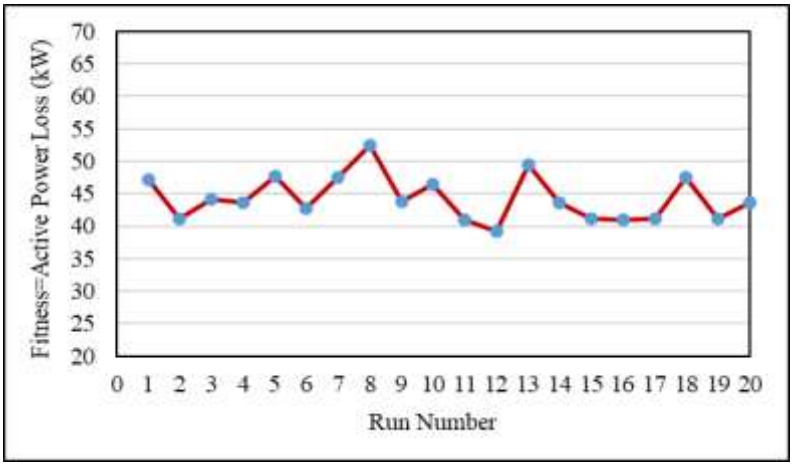



**Figure (7): GSA robustness test curve for IEEE 69 bus system**

The convergence performance of GSA for the IEEE 33-bus distribution network is illustrated in Figure 8, while Figure 9 presents the results for the IEEE 69-bus distribution network. The code was executed 20 times, and the best run was selected as the global solution. The global convergence performance was achieved with 300 iterations and a population size of 100.

The powerful of the presented optimization technique was compared with other work results as illustrated in Table 3 for the IEEE 33 bus system and in Table 4 for the IEEE 69 bus system. The presented optimization technique based on GSA provides results better than another algorithm.

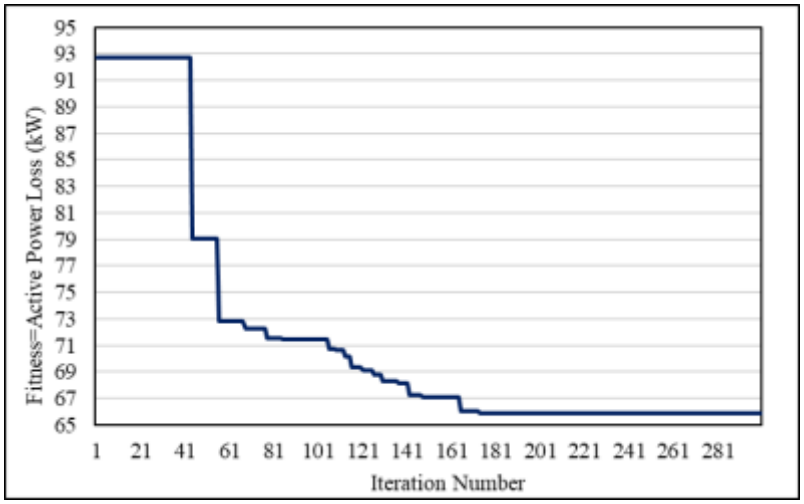



**Figure (8): GSA convergence performance curve for IEEE 33 bus system**

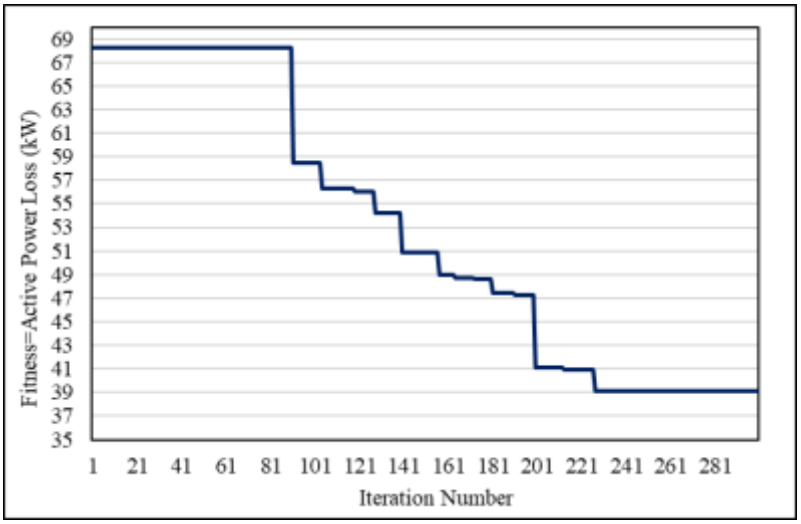



**Figure (9): GSA convergence performance curve for IEEE 69 bus system**

**Table 3: Simulation result comparison for IEEE 33 bus system**

| Reconfiguration and DG Sizing and Allocation | Open Switch | Total DG Output (MW) | Lowest Bus Voltage (pu) | Power Loss (kW) | Loss Reduction (%) |
|---|---|---|---|---|---|
| GA (Rao, et al., 2012) | 7, 10, 34, 28, 32 | 1.963 | .977 | 75.13 | 62.9 |
| RGA (Rao, et al., 2012) | 7, 9, 32, 12, 27 | 1.774 | .969 | 74.32 | 63.3 |
| HSA (Rao, et al., 2012) | 7, 32, 10, 14, 28 | 1.668 | .970 | 73.05 | 63.9 |
| FWA (Imran, et al., 2014) | 7, 1132, 14, 28, | 1.684 | .971 | 67.1 | 66.89 |
| EP (Ola Badran, et al., 2018) | 7, 8, 9, 28, 32 | 1.964 | .971 | 73.97 | 63.49 |
| PSO (Ola Badran, et al., 2018) | 7, 10, 13, 28, 32 | 1.766 | .974 | 72.42 | 64.3 |
| FA (Ola Badran, et al., 2018) | 7, 10, 13, 28, 32 | 1.825 | .975 | 72.36 | 64.28 |
| ISCA (Raut & Mishra, 2020) | 7, 14, 28. 31, 9 | 1.69120 | - | 66.81 | 67.03 |
| The Proposed Method by GSA | 32, 34, 27, 33, 9 | 1.822 | .9695 | 65.87 | 67.488 |

**Table 4: Simulation result comparison for IEEE 69 bus system**

| Reconfiguration and DG Sizing and Allocation | Open Switch | Total DG Output (MW) | Lowest Bus Voltage (pu) | Power Loss (kW) | Loss Reduction (%) |
|---|---|---|---|---|---|
| GA (Rao, et al., 2012) | 10, 45, 55, 62, 15 | 2.02920 | .97270 | 46.5 | 73.380 |
| GA (Rao, et al., 2012) | 10, 14, 55, 62, 16 | 2.06540 | .97420 | 44.230 | 8.320 |
| HSA (Rao, et al., 2012) | 69, 13, 58, 61, 17 | 1.87180 | .97360 | 4.3 | 82.080 |
| MPSO (Essallah & Khedher, 2020) | 14, 58, 61, 69, 70 | 2.2736 | .98994 | 42.2 | 81.1 |
| ISCA (Raut & Mishra, 2020) | 12, 9, 57, 63, 69 | 1.8731 | - | 39.73 | 82.34 |
| The Proposed Method by GSA | 69, 70, 71, 72, 73 | 2.429 | .9807 | 39.20 | 82.55 |

## 5. Conclusion

This paper presents an optimization methodology to simultaneously determine the optimal Distribution Network Reconfiguration (DNR), Distributed Generation (DG) sizing, and DG placement. The proposed approach aims to achieve the best voltage profile while minimizing active power loss in the distribution system. The Gravitational Search Algorithm (GSA) was employed to obtain the lowest fitness value. The effectiveness of the proposed technique was validated using a 33-bus distribution network, demonstrating its efficiency in simultaneously achieving optimal DNR, DG sizing, and DG placement. A portion of the obtained results was compared with existing published studies, demonstrating that the proposed optimization methodology achieved superior performance. Additionally, the results indicate that GSA outperforms other methods presented in previous research. Additionally, the limitations of the proposed methodology will appear in the case of dynamic load. The configuration of the network must change during the change of the load which may affect the network switches. This issue could be a future work. Authors should look forward to finding one configuration that is suitable during load change.

## Acknowledgments

"The author would like to thank Palestine Technical University-Kadoorie (PTUK) for supporting this research".

## References

- Abdelaziz, M. (2017). Distribution network reconfiguration using a genetic algorithm with varying population size. Electric Power Systems Research, 142, 9-11.
- Avchat, H. S., & Mhetre, S. (2020). Optimal placement of distributed generation in distribution network using particle swarm optimization. Paper presented at the 2020 International Conference for Emerging Technology (INCET).
- Badran, O. (2023). IEEE-69 Distribution Network Performance Improvement by Simultaneously Optimal Distributed Generation Sizing and Location Using PSO Algorithm.
- Badran, O., & Jallad, J. (2014). Experimental characterization of lead–acid storage batteries used in PV power systems.
- Badran, O., & Jallad, J. (2023a). Active and Reactive Power loss Minimization Along with Voltage profile Improvement for Distribution Reconfiguration. International journal of electrical and computer engineering systems, 14(10), 1193-1202.
- Badran, O., & Jallad, J. (2023b). Multi-Objective Decision Approach for Optimal Real-Time Switching Sequence of Network Reconfiguration Realizing Maximum Load Capacity. Energies, 16(19), 6779.
- Badran, O., & Jallad, J. (2023c). Multi-objective decision approach integrated with Loadability and weight factor analysis for reconfiguration with DG sizing and allocation including tap changer. Arabian Journal for Science and Engineering, 48(5), 6797-6818.
- Badran, O., Jallad, J., Mokhlis, H., & Mekhilef, S. (2020). Network reconfiguration and DG output including real time optimal switching sequence for system improvement. Australian Journal of Electrical and Electronics Engineering, 17(3), 157-172.

- Badran, O., Mekhilef, S., Mokhlis, H., & Dahalan, W. (2017). Optimal reconfiguration of distribution system connected with distributed generations: A review of different methodologies. Renewable and Sustainable Energy Reviews, 73, 854-867.
- Badran, O., Mokhlis, H., Mekhilef, S., & Dahalan, W. (2018). Multi-Objective network reconfiguration with optimal DG output using meta-heuristic search algorithms. Arabian Journal for Science and Engineering, 43, 2673-2686.
- Eldurssi, A. M., & O'Connell, R. M. (2014). A fast nondominated sorting guided genetic algorithm for multi-objective power distribution system reconfiguration problem. IEEE Transactions on Power Systems, 30(2), 593-601.
- Essallah, S., & Khedher, A. (2020). Optimization of distribution system operation by network reconfiguration and DG integration using MPSO algorithm. Renewable Energy Focus, 34, 37-46.
- Ha, M. P., Nazari-Heris, M., Mohammadi-Ivatloo, B., & Seyedi, H. (2020). A hybrid genetic particle swarm optimization for distributed generation allocation in power distribution networks. Energy, 209, 118218.
- Imran, A. M., Kowsalya, M., & Kothari, D. (2014). A novel integration technique for optimal network reconfiguration and distributed generation placement in power distribution networks. International Journal of Electrical Power & Energy Systems, 63, 461-472.
- Karunarathne, E., Pasupuleti, J., Ekanayake, J., & Almeida, D. (2021). The optimal placement and sizing of distributed generation in an active distribution network with several soft open points. Energies, 14(4), 1084.
- Karuppiah, N. (2021). Optimal siting and sizing of multiple type DGs for the performance enhancement of Distribution System using Differential Evolution Algorithm. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(2), 1140-1146.
- Kashem, M., Ganapathy, V., & Jasmon, G. (2000). Network reconfiguration for enhancement of voltage stability in distribution networks. IEE Proceedings-Generation, Transmission and Distribution, 147(3), 171-175.
- Mohandas, N., Balamurugan, R., & Lakshminarasimman, L. (2015). Optimal location and sizing of real power DG units to improve the voltage stability in the distribution system using ABC algorithm united with chaos. International Journal of Electrical Power & Energy Systems, 66, 41-52.
- Moradi, M. H., & Abedini, M. (2012). A combination of genetic algorithm and particle swarm optimization for optimal DG location and sizing in distribution systems. International Journal of Electrical Power & Energy Systems, 34(1), 66-74.
- Nguyen, T. T., & Truong, A. V. (2015). Distribution network reconfiguration for power loss minimization and voltage profile improvement using cuckoo search algorithm. International Journal of Electrical Power & Energy Systems, 68, 233-242.
- Pegado, R., Ñaupari, Z., Molina, Y., & Castillo, C. (2019). Radial distribution network reconfiguration for power losses reduction based on improved selective BPSO. Electric Power Systems Research, 169, 206-213.
- Rahim, M. N. A., Mokhlis, H., Bakar, A. H. A., Rahman, M. T., Badran, O., & Mansor, N. N. (2019). Protection coordination toward optimal network reconfiguration and DG sizing. IEEE Access, 7, 163700-163718.
- Rao, R. S., Ravindra, K., Satish, K., & Narasimham, S. (2012). Power loss minimization in distribution system using network reconfiguration in the presence of distributed generation. IEEE transactions on power systems, 28(1), 317-325.
- Raut, U., & Mishra, S. (2020). An improved sine–cosine algorithm for simultaneous network reconfiguration and DG allocation in power distribution systems. Applied Soft Computing, 92,106293.
- Sivanagaraju, S., Rao, J. V., & Raju, P. S. (2008). Discrete particle swarm optimization to network reconfiguration for loss reduction and load balancing. Electric power components and systems, 36(5), 513-524.
- Tyagi, A., Verma, A., & Bijwe, P. (2018). Reconfiguration for loadability limit enhancement of distribution systems. IET Generation , Transmission & Distribution, 12(1), 88-93.
- Yan, J., Shamim, T., Chou, S., Desideri, U., & Li, H. (2017). Clean, efficient and affordable energy for a sustainable future. Applied Energy, 185, 953-962.

Palestinian Journal for Open Learning & e-Learning No. (20) – May 2025
https://journals.qou.edu/index.php/jropenres
E-ISSN: 2074-5656 / P-ISSN: 2520-5692

# Feature Selection for Serving Medical Datasets Applying Heuristic Algorithms (Scatter Search within Decision Tree Classifier)

***Mr. Maher Ibrahim Issa****

lecturer , Computer Information Systems , Al-Quds University, Salfeet, Palestine.

**Oricd No**: 0009-0003-5742-9772
**Email**: missa@qou.edu

**Received:**
6/03/2024

**Revised:**
6/03/2024

**Accepted:**
7/04/2024

***Corresponding Author:**
missa@qou.edu

**Citation**:



**Abstract**

**Objectives**: This research presents a feature selection process on different datasets of the medical domain with different aims and sizes using a wrapper approach based on a powerful metaheuristic algorithm which is the Scatter Search Algorithm and J48 decision tree classifier as the selection criteria.

**Methods**: The paper applied a modified approach of the basic Sequential Scatter Search algorithm called Improved Sequential Scatter Search follows the basic procedures of the original algorithm in addition to an early improvement mechanism choosing decision tree classifier to be the evaluator of the experiments.

**Results**: The experimental results show competition and superiority in feature selection compared to other metaheuristic algorithms for the same datasets in consideration of number of features selected and accuracy.

**Conclusion**: This research emphasizes the importance of wrapper approaches using metaheuristic algorithms to select the most dominant attributes in a dataset which is very important in reduction of the cost and complexity of all data analysis areas.

**Keywords**: Metaheuristic (MH), feature selection (FS), scatter search Algorithm (SSA), decision tree (DT), medical datasets.

# اختيار الميزة لخدمة مجموعات البيانات الطبية من خلال تطبيق الخوارزميات الإرشادية (البحث المبعثر ضمن مصنف شجرة القرار)

***أ. ماهر ابراهيم عيسى****

[1]محاضر، انظمة معلومات حاسوبية، جامعة القدس المفتوحة، سلفيت، فلسطين.

**الملخص**

**الأهداف**: تعرض هذه الورقة عملية اختيار الميزات على مجموعات بيانات مختلفة في المجال الطبي بأهداف وأحجام مختلفة باستخدام نهج مجمّع يعتمد على خوارزمية ارشادية قوية وهي خوارزمية البحث المبعثرة ومصنف شجرة القرار J48 كمعيار للاختيار.

**الطرق**: طبقت هذه الورقة البحثية طريقة مستحدثة لخوارزمية البحث المبعثر اطلق عليها خوارزمية البحث المبعثر المحسن والذي اتبع خطوات الخوارزمية الرئيسية واضاف الية تحسين بتطبيق مصنف شجرة القرار كاداة تقييم للتجربة.

**النتائج**: اظهرت الدراسة ان النهج المستخدم اظهر تفوقا احيانا ومنافسة احيانا اخرى مقارنة مع خوارزميات ارشادية اخرى لنفس مجموعة البيانات على اساس معيارين هما: الدقة واختيار الميزة.

**في الختام**: يؤكد هذا البحث على اهمية النهج المجمع مع الخوارزميات الارشادية في اختيار الميزات السائدة من مجموعة البيانات والذي يعتبر مهما جدا في تقليل التكلفة والتعقيد في جميع مجالات تحليل البيانات.

**الكلمات المفتاحية**: الميتايورستك (MH)، اختيار الميزات (FS)، خوارزمية البحث المبعثر (SSA)،شجرة القرار (DT)، مجموعة البيانات الطبية.

## Introduction

Data mining and knowledge discovery in data (KDD) have been applied successfully in a number of study domains to extract new and useful knowledge from historical data (Ghazal and Hammad, 2022). Knowledge Discovery in Database (KDD) is growing at an unprecedented scale in medicine, industry, government, and civil society. Analysis and distilling knowledge from big data now drive many aspects of our society ،Insights gained from big data have revolutionized how we conduct business, governance, research, design, production, human interactions, and daily life (Shu and Ye, 2023).
High dimensionality seems to be the main challenge for implementation of the data mining techniques in different areas, in addition to noise, outliers, and errors in huge data sets. High-dimensional data presents a number of challenges for pattern identification. Additionally, smaller data sizes typically result in faster model training periods, which in turn speed research (Hancock et al., 2024). One popular technique for reducing data and comprehending feature information is feature selection (García-Pedrajas et al., 2021).

### Feature Selection

Feature Selection addresses the issue of high dimensionality. In order to get the best-performing subset of the original characteristics without any changes, it involves choosing the pertinent features and eliminating the redundant, noisy, and irrelevant ones (Bouchlaghem et al., 2022).
Reducing the dimensionality of the data is the initial stage in integrating low-dimensional data into a pattern recognition system. One of the most important tasks in pattern recognition research is creating an accurate system for recognizing patterns in high-dimensional data (Varma et al., 2022). When there are more dimensions than there are observations, this is referred to as high dimensionality. This makes computations extremely challenging. Data from high-dimensional space is reduced to a more manageable low-dimensional space using dimensionality reduction techniques (Probst and Reymond, 2020). Using feature selection, the influence of dimensionality on the dataset is minimized by identifying the subset of features that best captures the data (Abdulrazzaq and Saeed, 2019). It is helpful for identifying a good subset of features that is suited for the given problem since it extracts from the input data the significant and pertinent features for the mining process and eliminates redundant and irrelevant features (Ayesha et al., 2020). Creating a limited subset of features that accurately captures the essential aspects of the entire input data set is the primary goal of feature selection (Velliangiri and Alagumuthukrishnan, 2019). Feature selection reduces the amount of data, lowers the amount of storage required, improves prediction accuracy, prevents over fitting, and shortens the execution and training times for variables that are simple to grasp (Zebari et al., 2020). According to Dash and Liu (1997), in a typical attribute reduction method there are four basic steps (see Figure 1), i.e., (a) a generation procedure to generate the next candidate subset: (b) an evaluation function to evaluate the generated subset; (c) a stopping criterion to decide when to terminate the process; and (d) a validation procedure to check the validity of the subset (Dash and Liu, 1997).

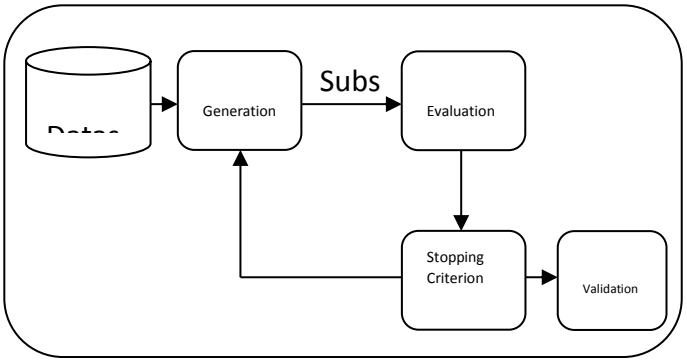



**Figure (1): Attribute reduction process with validation**

The goal of feature selection is to choose a subset of features based on redundancy and relevance from the original collection of features. Four categories of assessment approaches were first used in feature selection: filter, wrapper, embedding, and hybrid (Abd-Alsabour, 2018). Ensemble feature selection is a new kind of evaluation technique that has been created recently (Zebari et al., 2020). Here, the filter and wrapper models are discussed. Filter methods usually don't involve induction algorithm and evaluate the goodness of attributes subset cheaply using intrinsic characteristic of the data, while wrapper methods are computationally expensive as they used the induction algorithm to evaluate the attributes subset, but outperform filter methods in terms of predictive accuracy (Lyu et al., 2023).

Medical datasets are frequently utilized in data mining research in area of feature selection. A study by (Chen, C. W, 2020) introduced a combination of different types of feature selection algorithms, then results show that a combination of filter (i.e., principal component analysis) and wrapper (i.e., genetic algorithms) techniques by the union method is a better choice, providing relatively high classification accuracy. Nadimi-Shahraki (2021) used B-MFO in solving the feature selection problem for different medical datasets compared to other comparative algorithms and showed superiority in comparison to BDA, and BSSA algorithms.

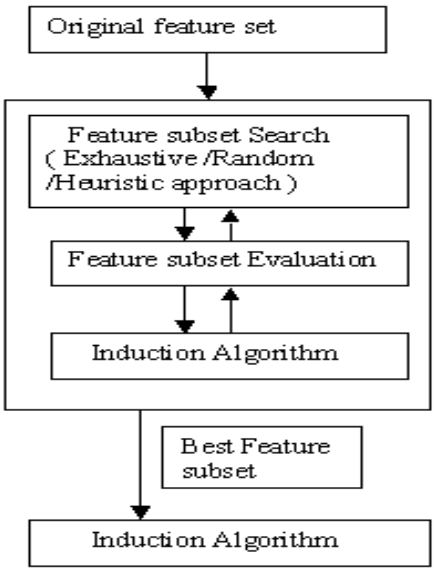
**Feature Subset Generation**

Feature selection as a search problem starts with feature generation. Feature subsets can be generated in a number of ways (Vandana and Chikkamannur, 2021; Hussain et al., 2021):

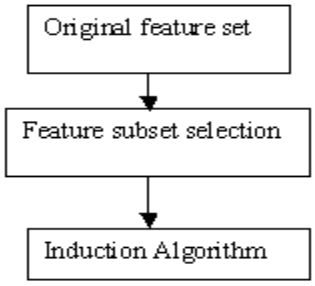
- Sequential Forward Selection (SFS) begins with the selection of the single best feature, which is determined by the objective function. The following stage involves creating a pair of features by combining the best feature and one feature from the remaining initial set. The pair with the best match is chosen. Using one of the remaining features, the next triplet of features is created, and the best triplet is chosen from the best pair that was previously chosen. Until a predetermined number of features are chosen, the process is repeated. When the optimal subset is tiny, it operates at its best. In SFS, a particular feature that might be helpful in the beginning might become unnecessary in later iterations for reasons that are not discoverable.
- Sequential Backward Selection (SBS) begins with an entire feature set as the beginning set. First, this starting collection is used to compute the criteria function. Every iteration involves the deletion of one feature, computation of the criterion feature for every subset, and deletion of the poorest feature. Until a certain number of features remain, this is repeated. It functions best when there is a big optimal subset. In SBS, a feature that is dropped in the first iteration might resurface in later iterations and be more helpful; however, this usefulness cannot be assessed.
- Bidirectional Search (BDS): This method applies SFS and SBS simultaneously to converge to the same answer, bringing a trade-off between the two.
- Random Generation is the fourth approach and performs the search process starting with randomly selected subset, then use sequential or bidirectional strategy.

Selection of some feature during the generation process is taken upon some evaluation measurement, such as classical ones: entropy, Bayesian, and Euclidean distance, or predictive classifier measure i.e. accuracy. The search process is iterative until reaching stopping criteria. The solution can be validated by univariate or multivariate approach. Decision Tree classifier is an example of univariate approaches, while neural network is a multivariate example.

The sequential search techniques in FS suffer from computational complexity since they are exhaustive procedures in their hunt for the best solution. (Elgamal et al., 2020). Current research focuses on meta-heuristic search, which, in contrast to optimum solutions, enables handling large-scale issue instances by providing a near-optimal solution in a reasonable amount of time (Sharma and Kaur, 2021).



a. Wrapper approach




b. Filter approach

**FIGURE (2) : WRAPPER VS. FILTER**

**Meta-heuristic Algorithms**

Metaheuristic algorithms have been categorized into two types based on solution search strategies (Sharma and Kaur, 2021): single-solution-based algorithms (S-Metaheuristic) and population-based algorithms (P-Metaheuristic). S-metaheuristics employ local search techniques to enhance a single solution iteratively until a stopping condition is met. Examples of S-metaheuristic algorithms include Simulated Annealing (SA), Hill Climbing (HC), Record-to-Record (RR), and Threshold Accepting. On the other hand, P-metaheuristics, such as Genetic Algorithm (GA) and Scatter Search (SS), utilize evolutionary approaches to improve a population of solutions iteratively. Additionally, algorithms like Ant Colony Optimization, Bee Colony Optimization, and Bat Algorithm are also categorized as P-metaheuristic, drawing inspiration from natural systems (Jaddi and Abdullah, 2020). P-metaheuristics evolve a population of solutions iteratively until a predetermined stopping criterion is satisfied.

Despite the early idea of Scatter Search (SS), the FS problems implementation for this algorithm is still at the beginning. Few studies such as (Wang et al., 2009; Wang et al., 2012; López et al., 2006) handle FS using SSA in different manner. (Wang et al., 2009) Used SSA to enhance the attribute reduction process in rough set theory. The approach of this study applied at different datasets and compared the result to other Computational Intelligence (CI) algorithms tried to address the same problem. The result shows promising and competitive performance in terms of solution quality, and superior performance in terms of computational costs. The other study (Wang et al., 2012) proposed a novel approach based on rough set and SSA, invoking entropy for searching optimal solution. The approach has been applied for two international credit scoring datasets and shows huge saving in computational costs and higher accuracy compared to base classification methods. The previous two studies based on MATLAB coding. López et al. (2006) applies FS using SSA on different 8 datasets with parallel computers for two combination methods, and compared results to basic SSA and Genetic algorithm supported by WEKA tool.

This paper introduces a novel wrapper feature selection methodology that combines the Scatter Search Algorithm (SSA) with the Decision Tree (DT) J48 classifier for application on standard medical datasets. The proposed approach generates an initial set of solutions termed as the RefSet, which undergoes refinement through a local search procedure. Subsequently, these solutions are evaluated using DT to identify the optimal solution. The chosen solution then undergoes further enhancement. This iterative process continues until a predefined stopping criterion is met, typically defined as the best solution in terms of both accuracy and the number of features selected. The approach is systematically applied across various parameters. The primary innovation lies in the inclusion of an early improvement mechanism following the creation of initial solutions, a step not typically addressed in traditional SSA approaches. The improvement strategy is explained later in a separate section.

**Decision tree classifier**

Decision Tree (DT) is a widespread classification algorithm that partitions the data using information gain until all instances reach uniform class labels (Huan and Hiroshi, 1998). Each time a single feature used as splitting node according to its values, and information gain used to determine the split feature F. DT achieves high accuracy level of prediction despite high computational cost (Wang et al., 2009).

Data in DT approach is divided into two datasets (Tan et al., 2006): Training set (large) and Test set (small) before starting the induction. Basic Induction Algorithm of DT is as follows:

- Initialize by setting variable T to be the training set
- Apply the following steps to T:

1. If all elements in T are of class X, create node X and halt.
2. Otherwise, select a feature F with values v1, v2…, vn, and partition T into T1, T2…, Tn according to their values in F, then create a parent node with F and T1, T2…, Tn as child nodes
3. Apply the procedure recursively to each child node.

One of the robust algorithms for decision tree (DT) induction is the C4.5 algorithms. It represents an advancement over classical divide-and-conquer approaches (Bahar and Saad, 2024). C4.5 utilizes gain ratio as the basis for splitting the training set and incorporates methods to handle missing values, noisy data, and numeric attributes (Cherfi et al., 2020; Quinlan, 1996; Sugumaran et al., 2007). In our computational analysis, we employed the J48 decision tree, which implements the C4.5 criteria and is provided by the WEKA software suite.

**Scatter search algorithm**

The foundation of Scatter Search (SS) was introduced by Glover in 1977 as an evolutionary approach for addressing combinatorial and nonlinear optimization problems. SS follows a systematic problem-solving approach outlined in the following steps (Ghamisi and Benediktsson, 2014):

1. Generate starting set of solution vectors by heuristic processes, then choose a subset of the best vectors to be reference solutions.
2. Perform linear combinations of the subsets of the reference solutions and create new solutions.
3. Extract the best solutions in procedure 2, and use it as starting solutions.
4. Repeat steps until specified iteration limit.

It is good to know that heuristic processes in SS are not of uniform design but represent a varied collection of procedures, also SS can start with infeasible solution to reach elite ones. SSA can be reflected to solve FS problems following the same way. Figure (3) shows the pseudo code of the Sequential Scatter Search (SSS) as the improvement mechanism executed sequentially (Garcıa-López et al., 2003).

```
            procedure Sequential Scatter Search
begin
    CreatePopulation(Pop,PopSize);
    GenerateReferenceSet(RefSet, RefSetSize);
     repeat
            repeat
                SelectSubset(Subset, SubSetSize);
                        CombineSolutions(SubSet, CurSol);
                ImproveSolutions(CurSol, ImpSol);
           until (StoppingCriterion 1);
           UpdateReferenceSet(RefSet);
   until (StoppingCriterion2);
end
```

FIGURE (3): Sequential Scatter Search Algorithm

**Solution Representation**

Scatter Search Algorithm uses binary representation for solutions. The solution is a ‘0’,’1’ vector (Ghamisi and Benediktsson, 2014), its size equal to the number of conditional attributes in a dataset. If a feature is selected in this subset the corresponding index will have 1, while 0 indicates that the feature is not selected in this solution.

**Population Generation**

In this step a set of initial solutions vectors are generated. These solutions are divided as diverse and good quality solutions. In our approach we use the vector of weights used by (López et al., 2006) to generate a solution. This strategy considers the vector of weights of features P(X) = P(X1),.,P(Xc) where C is the number of conditional features, and given by P(Xi) = ft({Xi}) where these weights indicate the quality of the features for classifying by itself. Let L be the set of features Xi with the highest weight P(Xi), then select randomly good feature (high P) from L where the presence of this feature improves the set.

**Reference Set Generation**

Two sets of solutions are added to each other to form the reference set (RefSet). The first one (RefSet1) is good solutions set, while the other (RefSet2) is the set of diverse solutions. In order to achieve diversity, the symmetric difference criteria had been used with the procedures explained as follow: Let S be any solution and C is the set of features belong to any solution in the RefSet, then the diversity of each solution S is given by the following equation:

Div (S) = Diff (S,C) = |(SUC)/ (S∩C)|

For generating the RefSet, SS usually considers all subsets of two solutions in the current set of solutions to be combined to generate new solutions.

**Solutions Combination**

This procedure generates new solutions by combining subsets of the original reference solution. In this approach the suggested greedy combination (GC) method by (López et al., 2006) had been used. Let S1 and S2 be solutions in the subset. GC method generates two new solutions S1', and S2'. First, GC starts by adding the common features of S1 and S2 to the new solutions S1' and S2'. Then at each iteration one of the remaining features in S1 or S2 is added to S1' or S2'.

**Improvement Method**

This method is applied to each solution obtained by the described combination strategy. The purpose of this method is to add some characteristics to the solution that improve it. The solution S accepts to add a feature if this improves its quality. If the solution has not been enhanced, then the output solution is the same as the input solution.

**Updating Refrence Set**

The reference set is updated after the previous procedures to be formed from two parts, the first one is the RefSet1 obtained by improvement method, where the second one RefSet2 is the produced by the diversity criteria explained earlier.

**Medical Datasets Description**

The Medical Datasets are widely used in data mining and machine learning researches for serving health care area. In our approach we chose to apply our approach on 7 well-known medical datasets and compared to a research paper that found the best feature selection results for these datasets in last 14 years in terms of accuracy and selected features. Using these medical datasets is justified by their frequent utilization in data mining research, facilitating easy comparison of results with best prior study. Mullins et al. (2006) provided a concise overview of these datasets, which are briefly described below.

1. **Breast-Cancer**: This dataset typically contains features derived from digitized images of breast cancer biopsies. These features can include characteristics like cell nuclei properties and are used to classify whether a tumor is benign or malignant.
2. **Breast-w**: This dataset is also related to breast cancer diagnosis. It may contain different features or be derived from a different source, but its purpose is likely similar: to predict whether a tumor is benign or malignant.
3. **Heart-c:** This dataset is often used for predicting the presence of heart disease. It typically includes various clinical parameters such as blood pressure, cholesterol levels, and electrocardiogram readings, among others.
4. **Heart-stolog:** Another dataset related to heart disease diagnosis. It might contain similar features to the "Heart-c" dataset but could be from a different source or have variations in the features included.
5. **Hepatitis**: This dataset involves data related to the diagnosis and treatment of hepatitis, a liver inflammation caused by viral infection. It may include various clinical and laboratory parameters for patients with hepatitis.
6. **Lung-Cancer:** This dataset pertains to the diagnosis and prognosis of lung cancer. It might include features such as tumor size, histological type, and patient demographics, among others.
7. **Dermatology:** This dataset typically includes various attributes related to dermatological conditions, such as symptoms, patient characteristics, and diagnostic results. It's often used for predicting skin diseases or analyzing patterns in dermatological data.

**Table (1): Datasets description**

| # | Data Set | No. of Features | No. of Objects |
|---|---|---|---|
| 1 | Breast-cancer | 9 | 286 |
| 2 | Breast-w | 9 | 699 |

| # | Data Set | No. of Features | No. of Objects |
|---|---|---|---|
| 3 | Heart-c | 13 | 303 |
| 4 | Heart-stolog | 13 | 270 |
| 5 | Hepatitis | 19 | 155 |
| 6 | Lung-Cancer | 56 | 32 |
| 7 | Dermatology | 34 | 366 |

**The proposed approach**

The Improved Sequential Scatter Search (ISSS) builds upon the foundation of the basic Scatter Search (SSS) method described previously. The key distinction lies is the integration of an early improvement technique following the generation of initial solutions, illustrated in Figure 4. This essentially constitutes a double improvement approach. Consequently, the reference set formed comprises superior solutions, enhancing the potential for generating improved solutions.

```
procedure Improved Sequential Scatter Search
begin
    CreatePopulation(Pop,Popsize);
    ImproveInitialPop;
    GenerateReferenceSet(RefSet, RefSetSize);
    repeat
        repeat
            SelectSubset(Subset, SubSetSize);
            CombineSolutions(Subset, Cursol);
            ImproveSolutions(CurSol, ImpSol);
        Until (StoppingCriterion1);
        UpdateReferenceSet(RefSet);
    Until (StoppingCriterion2);
End
```

**FIGURE (4): ISSS pseudo code**

In a more expressive manner, the following illustrations Figure 5 and Figure 6 clarify the difference between the basic Sequential Scatter Search and the Improved one which is our approach. The improvement step includes an accuracy evaluation mechanism using Decision Tree. The early improvement step helps to generate an initial strong Reference Set.

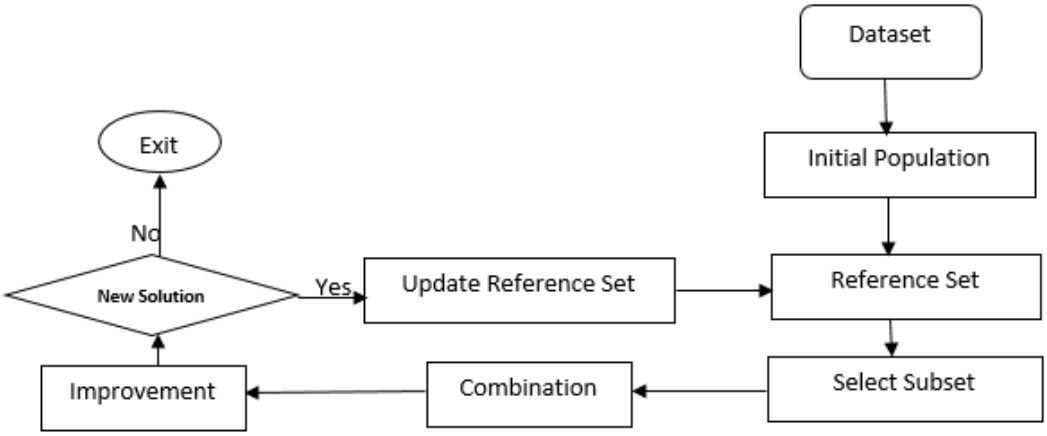



**FIGURE (5): Basic SSS**

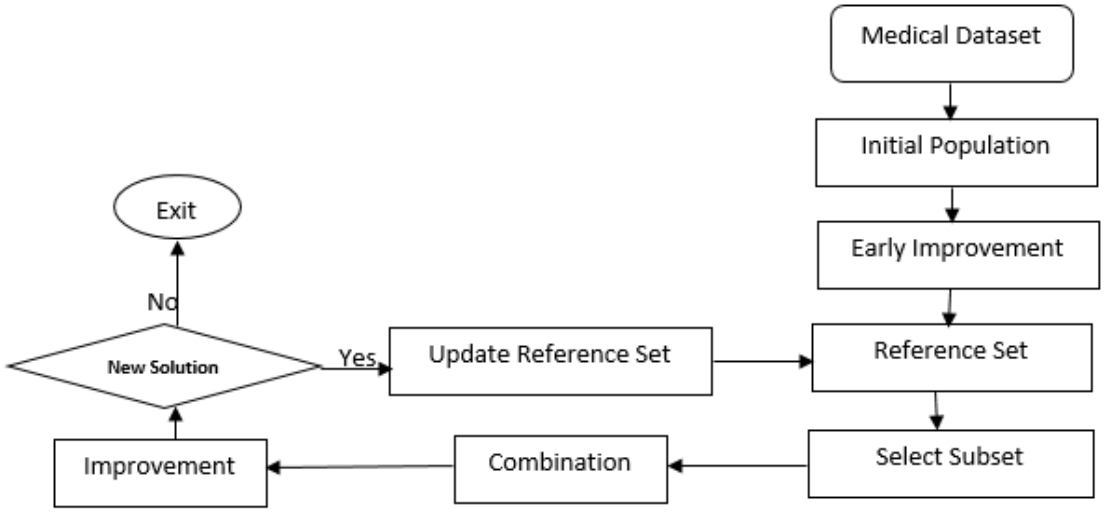



**FIGURE (6): Improved ISSS**

**Experiment and results**

The computational experiments performed for this paper aims mainly to show the performance of our wrapper approach (Scatter Search Algorithm and J48 classifier evaluator) in feature selection for serving medical datasets, and the competitively with other MH algorithms. The datasets were taken from UCI repository in (ARFF) format to fit WEKA requirements. The chosen medical datasets are of different branches and different sizes, and most of them already have been used in previous FS studies. Datasets are shown in table 1 above. All experiments had been carried out using a Portable Computer Duo Core CPU with 2.2 GHz speed, and has 4GB memory. We applied the approach on all datasets using fixed and variable parameters.

**ISSS Experiments**

In this part the size the feature selection process was performed with three different population size which is (|C|/2, |C|, 2|C|) to see the best result produced according to features and accuracy. The accuracy had been calculated using 10-fold cross validation. Table (2) show the experiment results for 5 or more runs without any change in results.

**Table (2): No. of selected features and avg. accuracy for different population size**

| | PopSize=\|C\|/2 | | PopSize=\|C\| | | PopSize=2\|C\| | |
|---|---|---|---|---|---|---|
| Data Set | Selected Features By ISSS | Avg.Accuracy % | Selected Features By ISSS | Avg.Accuracy % | Selected Features By ISSS | Avg.Accuracy % |
| Breast-cancer | 2 | 74.8 | 2 | 74.8 | 3 | 74.5 |
| Breast-w | 3 | 95.4 | 3 | 95.6 | 3 | 95.6 |
| Heart-c | 3 | 81.5 | 4 | 83.5 | 5 | 82.5 |
| Heart-stolog | 3 | 78.5 | 3 | 85.2 | 3 | 85.2 |
| Hepatitis | 2 | 85.1 | 1 | 84.5 | 2 | 85.1 |
| Lung-Cancer | 2 | 68.8 | 2 | 75 | No result | No result |
| Dermatology | 8 | 92.3 | 8 | 92.3 | No result | No result |

From the information given in Table (2), the effect of using different size of population size can be noticed. In average using bigger population size increases the strength of the result according to accuracy and features. But it is obvious that

using big population size leads us to unacceptable computational time. So the best results according to our experiment for ISSS approach happens when PopSize=|C|.

In order to recognize the strength of the proposed approach, the results should be compared to other approached. In the study conducted by Kilic, Essiz, and Keles (2023), it was noted that by Palanisamy and Kanmani (2012) achieved the best results by using ABC (Artificial Bee Colony) for feature selection for the same medical datasets used in this paper. However, they tried to improve results using binary ASO (Anarchic Society Optimization) algorithm but their approach has superiority just in Heart-C Dataset as the number of features selected was 5 with the same accuracy.

The following table3 is a comparison between our approach results when PopSize=|C| and the results mentioned by Palanisamy and Kanmani (2012) which is already has been compared to other approaches and shows superiority. It's noteworthy to refer that ABC stands for Artificial Bee Colony Algorithm used for Feature Selection.

**Table (3): Results comparison between ISSS and ABC**

| Dataset | FS by ISSS | accuracy | FS by ABC | accuracy |
|---|---|---|---|---|
| Heart-C | 2 | 74.8 | 6-7 | 86.92 |
| Dermatology | 8 | 92.3 | 24 | 98.55 |
| Hepatitis | 1 | 84.5 | 11 | 81.26 |
| Lung Cancer | 2 | 75 | 27 | 89.25 |
| Breast-w | 3 | 95.6 | 4 | 96.99 |
| Heart-stolog | 3 | 85.2 | 6 | 84.07 |

In Table (3), the ISSS approach was implemented with classification accuracy (CA) as the fitness function. The results demonstrate the power and competitiveness of our approach in terms of both the number of selected features and CA. For certain datasets (e.g., heart-c, lung), ISSS excels in choosing fewer features but may yield lower CA. However, for others (e.g., heart-stolog, hepatitis), ISSS demonstrates superiority in both feature selection and accuracy.

Obviously, the wrapper-based approach of ISSS proves its superiority over the filter-based approach in terms of the number of selected attributes. However, despite its efficacy, ISSS does have certain drawbacks. Chief among them is its computational time, particularly when dealing with large datasets. The method demands significant computational resources and can sometimes reach unacceptable levels of time.

It is important to mention that Weka software supports several original algorithms, including Naive Bayes and Support Vector Machine. We applied Naive Bayes to these datasets, but the results were poor compared to our approach and the other studies discussed here.

## Conclusion and future work

Scatter Search proves to be a promising approach for feature selection problems. In the medical domain, the ISSS wrapper method effectively conducts feature selection, leveraging its built-in diversification and intensification mechanisms. This method demonstrates clear superiority in both the quantity of selected features and competitiveness in accuracy.

For future studies, we suggest integrating new diversification techniques and improvement methods to enhance the performance and accuracy of solutions, catering to diverse domain datasets. Additionally, exploring different algorithms and approaches for feature selection in medical datasets holds significant potential for advancing research in this area

## REFERENCES


- Abd-Alsabour, N. (2018). On the Role of Dimensionality Reduction. J. Comput., 13(5), 571-579.
- Abdulrazzaq, M. B., & Saeed, J. N. (2019, April). A comparison of three classification algorithms for handwritten digit recognition. In 2019 International Conference on Advanced Science and Engineering (ICOASE) (pp. 58-63). IEEE.
- Ayesha, S., Hanif, M. K., & Talib, R. (2020). Overview and comparative study of dimensionality reduction techniques for high dimensional data. Information Fusion, 59, 44-58.
- Bahar, M. H., & Saad, H. (2024). Decision Tree Induction Using Evolutionary Algorithms: A Survey. International Journal of Computing and Digital Systems, 15(1), 99–113. https://doi.org/10.12785/ijcds/150109
- Bouchlaghem, Y., Akhiat, Y., & Amjad, S. (2022). Feature Selection: A Review and Comparative Study. E3S Web of Conferences, 351, 01046. https://doi.org/10.1051/e3sconf/202235101046
- Chen, C. W., Tsai, Y. H., Chang, F. R., & Lin, W. C. (2020). Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. Expert Systems, 37(5), e12553.

- Chen, J., Yuan, S., Dongdong Lv, & Xiang, Y. (2021). A novel self-learning feature selection approach based on feature attributions. Expert Systems with Applications, 183, 115219–115219. https://doi.org/10.1016/j.eswa.2021.115219
- Cherfi, A., Nouira, K., & Ferchichi, A. (2020). MC4.5 decision tree algorithm: an improved use of continuous attributes. International Journal of Computational Intelligence Studies, 9(1/2), 4. https://doi.org/10.1504/ijcistudies.2020.106485
- Dash, M., & Liu, H. (1997). Feature selection for classification. Intelligent data analysis, 1(1-4), 131-156.
- Elgamal, Z. M., Yasin, N. B. M., Tubishat, M., Alswaitti, M., & Mirjalili, S. (2020). An improved Harris hawks optimization algorithm with simulated annealing for feature selection in the medical field. IEEE access, 8, 186638-186652.
- Erfani, S. M., Rajasegarar, S., Karunasekera, S., & Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class.
- Garcıa-López, F., Melián-Batista, B., Moreno-Pérez, J. A., & Moreno-Vega, J. M. (2003). Parallelization of the scatter search for the p-median problem. Parallel computing, 29(5), 575-589.
- García-Pedrajas, N., del Castillo, J. A. R., & Cerruela-García, G. (2021). SI(FS)2: Fast simultaneous instance and feature selection for datasets with many features. Pattern Recognition, 111, 107723. https://doi.org/10.1016/j.patcog.2
- Ghamisi, P., & Benediktsson, J. A. (2014). Feature selection based on hybridization of genetic algorithm and particle swarm optimization. IEEE Geoscience and remote sensing letters, 12(2), 309-313.
- Ghazal, M. M., & Hammad, A. (2022). Application of knowledge discovery in database (KDD) techniques in cost overrun of construction projects. International Journal of Construction Management, 22(9), 1632-1646.
- Glover, F., Laguna, M., & Martí, R. (2003). Scatter search. Advances in evolutionary computing: theory and applications, 519-537.
- Hancock, J. T., Wang, H., Khoshgoftaar, T. M., & Liang, Q. (2024). Data reduction techniques for highly imbalanced medicare Big Data. Journal of Big Data, 11(1), 8.
- Hussain, K., Neggaz, N., Zhu, W., & Houssein, E. H. (2021). An efficient hybrid sine-cosine Harris hawks optimization for low and high-dimensional feature selection. Expert Systems with Applications, 176, 114778.
- Jaddi, N. S., & Abdullah, S. (2020). Global search in single-solution-based metaheuristics. Data Technologies and Applications, 54(3), 275–296. https://doi.org/10.1108/dta-07-2019-0115
- Kaur, N., Singla, J., Mathur, G., Talwani, S., & Malik, N. (2023, November). An Advanced Feature Selection Approach to Improve Intrusion Detection System using Machine Learning. In 2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 984-992). IEEE.
- KILIC, U., ESSIZ, E. S., & KELES, M. K. (2023). Binary anarchic society optimization for feature selection. Romanian J. Inf. Sci. Technol, 26, 351-364.
- López, F. G., Torres, M. G., Batista, B. M., Pérez, J. A. M., & Moreno-Vega, J. M. (2006). Solving feature subset selection problem by a parallel scatter search. European Journal of Operational Research, 169(2), 477-489.
- Lyu, Y., Feng, Y., & Sakurai, K. (2023). A survey on feature selection techniques based on filtering methods for cyber-attack detection. Information, 14(3), 191.
- Mullins, I. M., Siadaty, M. S., Lyman, J., Scully, K., Garrett, C. T., Miller, W. G., ... & Knaus, W. A. (2006). Data mining and clinical data repositories: Insights from a 667,000 patient data set. Computers in Biology and Medicine, 36(12), 1351-1377.
- Nadimi-Shahraki, M. H., Banaie-Dezfouli, M., Zamani, H., Taghian, S., & Mirjalili, S. (2021). B-MFO: a binary moth-flame optimization for feature selection from medical datasets. Computers, 10(11), 136.
- Palanisamy, S., & Kanmani, S. (2012). Artificial bee colony approach for optimizing feature selection. International Journal of Computer Science Issues (IJCSI), 9(3), 432.
- Probst, D., & Reymond, J. L. (2020). Visualization of very large high-dimensional data sets as minimum spanning trees. Journal of Cheminformatics, 12(1), 12.
- Quinlan, J. R. (1996). Improved use of continuous attributes in C4. 5. Journal of artificial intelligence research, 4, 77-90.
- Sharma, M., & Kaur, P. (2021). A comprehensive analysis of nature-inspired meta-heuristic techniques for feature selection problem. Archives of Computational Methods in Engineering, 28, 1103-1127.
- Shu, X., & Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. Social Science Research, 110, 102817.
- Sugumaran, V., Muralidharan, V., & Ramachandran, K. I. (2007). Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing. Mechanical systems and signal processing, 21(2), 930-942.
- Tan, P.-N., Steinbach, M., & Pearson, V. (2006). Introduction to Data Mining.WP CO.

- Vandana, C. P., & Chikkamannur, A. A. (2021). Feature selection: An empirical study. International Journal of Engineering Trends and Technology, 69(2), 165-170.
- Varma, K., Ajmire, P. E., & Rehapande, A. B. (2022). A REVIEW OF DIMENSIONALITY REDUCTION TECHNIQUES FOR HIGH DIMENSIONAL DATA. Journal of the Oriental Institute, 71(4).
- Velliangiri, S., & Alagumuthukrishnan, S. J. P. C. S. (2019). A review of dimensionality reduction techniques for efficient computation. Procedia Computer Science, 165, 104-111.
- Wang, J., Hedar, A. R., Wang, S., & Ma, J. (2012). Rough set and scatter search metaheuristic based feature selection for credit scoring. Expert Systems with Applications, 39(6), 6123-6128.
- Wang, J., Hedar, A. R., Zheng, G., & Wang, S. (2009, April). Scatter search for rough set attribute reduction. In 2009 International Joint Conference on Computational Sciences and Optimization (Vol. 1, pp. 531-535). IEEE.
- WEKA: A Java Machine Learning Package, https://ml.cms.waikato.ac.nz/weka/
- Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. Journal of Applied Science and Technology Trends, 1(1), 56-70.

**Palestinian Journal of Technology and Applied Sciences No – (8). June 2025**
**https://journals.qou.edu/index.php/PJTAS**
**E-ISSN: 2521-411X/ P-ISSN: 2520-7431**

# Using Fine Needle Aspiration Data to Classify Breast Cancer Types by Machine Learning

***Mr. Rami Suleiman Khader [1]*, Dr. Mohamed Mahmoud Dweib[2], Prof. Yousef Saleh Abuzir [3]***

1Master's Student, Faculty of Applied Science and Technology, Al-Quds Open University, Jenin, Palestine
**Oricd No**: 0009-0000-6520-8134
**Email**: rami_s_khader@msn.com

2 Associate Professor of Computer Systems, Faculty of Technology and Applied Sciences, Al-Quds Open University, Bethlehem, Palestine
**Oricd No**: 0000-0001-7493-9780
**Email**: mdweib@qou.edu

3Professor of Computer Systems, Faculty of Technology and Applied Sciences, Al-Quds Open University, Salfit, Palestine
**Oricd No**: 0000-0002-1220-1411
**Email**: yabuzir@qou.edu

**Received:**
7/03/2024

**Revised:**
7/03/2024

**Accepted:**
26/03/2024

***Corresponding Author**:
rami_s_khader@msn.com

**Citation**: Khader , R. S., Dweib, M. M., & Abuzir , Y. S. Using Fine Needle Aspiration Data to Classify Breast Cancer Types by Machine Learning. Palestinian Journal of Technology and Applied Sciences (PJTAS), 1(8). https://doi.org/10.33977/2106-000-008-003



- **Open Access**



## Abstract

**Objectives**: Breast cancer, a leading cause of death worldwide and the foremost in Palestine, often benefits from early diagnosis to improve patient outcomes. However, diagnosing small tumors accurately can be challenging, with a high risk of human error. This study seeks to enhance breast cancer classification by utilizing machine learning (ML) algorithms.

**Methods**: The research analyzed and utilized three machine learning techniques - Decision Tree Classifier (DTC), Support Vector Machine (SVM), and Random Forest Classifier (RFC) - to predict breast cancer tumors. The accuracy of the three algorithms was analyzed and evaluated using a confusion matrix as well as different metrics on a dataset containing 569 samples and 29 features.

**Results**: The result showed that the Decision Tree Classifier (DTC) has the high scores of 100% in accuracy, precision, sensitivity, and specificity.

**Conclusions**: In the conclusion, the research emphasizes the excellent performance of the Decision Tree Classifier in classifying breast cancer, which could significantly improve diagnostic accuracy and patient outcomes. The results indicate that DTC has the potential to be a useful ML model in decreasing human diagnostic mistakes and enhancing the early detection and care in medical environments, prompting additional studies to enhance and confirm its effectiveness.

**Keywords**: Machine Learning (ML), Breast Cancer Classifications, Decision Tree Classifier (DTC), Support Vector Machine (SVM), Random Forest Classifier (RFC), and Fine Needle Aspiration.

# استخدام بيانات الخزعة المسحوبة لتصنيف أنواع سرطان الثدي عن طريق التعلم الآلي

***أ. رامي سليمان خضر [1]*، د. محمد محمود ذويب[2]، أ.د. يوسف صالح أبو زر [3]***

[1]طالب ماجستير، كلية التكنولوجيا والعلوم التطبيقية، جامعة القدس المفتوحة، جنين، فلسطين.

[2]أستاذ مشارك نظم الحاسوب، كلية التكنولوجيا والعلوم التطبيقية، جامعة القدس المفتوحة، بيت لحم، فلسطين.

[3]أستاذ دكتور نظم الحاسوب، كلية التكنولوجيا والعلوم التطبيقية، جامعة القدس المفتوحة، سلفيت، فلسطين.

## الملخص

**الاهداف:** يعد سرطان الثدي السبب الرئيسي للوفاة في جميع أنحاء العالم والأهم في فلسطين، يستفيد غالبا من التشخيص المبكر لتحسين نتائج المرضى. ومع ذلك، فإن تشخيص الأورام الصغيرة بدقة يمكن أن يكون صعبًا، مع ارتفاع مخاطر الخطأ البشري. تهدف هذه الدراسة إلى تعزيز تصنيف سرطان الثدي من خلال الاستفادة من خوارزميات التعلم الآلي.

**المنهجية:** قام البحث بتحليل ومقارنة ثلاث تقنيات للتعلم الآلي – مصنف شجرة القرار (DTC) وآلة المتجهات الداعمة (SVM) ومصنف الغابة العشوائية - (RFC) لتحديد الطريقة الأكثر كفاءة لتصنيف أورام سرطان الثدي. تم تقييم دقة الخوارزميات باستخدام مصفوفة الارتباك على مجموعة بيانات تحتوي على 569 عينة و29 ميزة.

**النتائج:** أظهرت النتائج أن مصنف شجرة القرار (DTC) كان الأكثر نجاحًا، حيث حقق درجات خالية من العيوب بنسبة 100٪ في الدقة والإحكام والحساسية والخصوصية.

**الخلاصة:** وفي الختام، يؤكد البحث على الأداء الممتاز لمصنف شجرة القرار في تصنيف سرطان الثدي، مما قد يحسن بشكل كبير من دقة التشخيص ونتائج المرضى. تشير النتائج إلى أن التشخيص المباشر للمصابين بالسرطان لديه القدرة على أن يكون أداة مفيدة في تقليل الأخطاء التشخيصية وتعزيز التعرف المبكر والرعاية في البيئات الطبية، مما يدفع إلى إجراء دراسات إضافية لتعزيز وتأكيد فعاليته.

**الكلمات المفتاحية:** التعلم الآلي (ML)، تصنيفات سرطان الثدي، مصنف شجرة القرار (DTC)، آلة الدعم المتجه (SVM)، مصنف الغابة العشوائية (RFC)، وسحب الخزعة.

## Introduction

In recent years, the integration of machine learning techniques into medical diagnostics has revolutionized the way diseases, particularly cancer, are classified and treated.

According to the World Health Organization (WHO, 2024), cancer is the second deadliest disease globally, causing approximately 9.6 million deaths annually. Among cancers, breast cancer ranks as the second most deadly globally. In Palestine, specifically the West Bank, 3,191 cancer cases were reported in 2021 (MHPS, 2021). Of these, breast cancer was the most common, with 526 cases, representing 16.5% of all cancers (UCI, 1995). This highlights the significant impact of breast cancer on the population in Palestine and the need for increased awareness and resources for prevention and treatment.

In medical terms, there are two types of tumors: benign and malignant. "Benign masses generally have a low density with well-defined margins and a fat covering over the lesion; whereas malignant masses generally have a slightly irregular shape, without symmetry and do not have fat .(CDC, 24) Fine Needle Aspiration (FNA) (ENT 2024) is one type of biopsy used to specify the cancer type, but there is a possibility of human error, especially when the tumor is small. For this reason, machine learning is preferred to be used. High danger disease, which affects humanity, is BC. Mainly, there is two types of this disease (Malignant and Benign) (Li, S., & Margolies, L. R. 2019), in some cases there are human mistakes to distinguish between these types. The research focuses on how to classify between these types based on Fine Needle Aspiration (FNA) metadata. The rationale for selecting this biopsy is available anywhere and cheap, and there are a lot of datasets available if researchers want to defend their results. Breast cancer is characterized by the uncontrolled growth of cells in the breast, with its specific type determined by the affected cell types (Rokach & Maimon, 2008; Juanjuan & Bradley, 2021). To diagnose breast cancer, fine needle aspiration (FNA) is commonly used. This biopsy technique involves inserting a thin needle into abnormal tissue or fluid for examination. FNA is generally considered a safe procedure, with complications occurring infrequently (Maglogiannis et al., 2009).

The presence of current machine learning models capable of classifying breast cancer does not negate the need for the development of new models. This requirement is propelled by the existing models' potential limitations in accurately detecting different cancer subtypes and early stages. New models can be deliberately crafted to confront these unique challenges, potentially enhancing overall performance in breast cancer classification. As a result of that, using machine learning technologies into public health, including medical diagnosis, which has revolutionized the way diseases are classified and treated, especially cancer.

Despite this progress in using machine learning to classify breast cancer, some current models can face problems or challenges such as inaccuracy, limited generalization ability, and difficulties in distinguishing between cancer subtypes. These challenges arise due to factors such as dataset variability, feature noise, and model overfitting.

This study aims to develop novel ML models capable of superior accuracy and consistency in breast cancer classification across diverse datasets. To achieve this, it will:

- Build and evaluate novel ML models for breast cancer detection.
- Dat preprocessing and feature engineering to enhance and improve model accuracy.
- Develop general models capable of dealing with diverse datasets.
- Compare new models to existing standards using different performance metrics.
- Help minimize human diagnostic errors in breast cancer, particularly for complex cases.

The motivation for this research stems from the critical need for improved breast cancer diagnostic tools, particularly in regions with high prevalence rates like Palestine. By addressing the limitations of current models and leveraging recent advancements in ML, this study seeks to enhance patient outcomes and early detection.

**This research contributes to the field by:**

- Introducing novel ML models for breast cancer classification.
- Improving data preprocessing phases and feature engineering process.
- Performing evaluation of capabilities and performance for the three ML model.
- Emphasizing the potential of these ML models to improve breast cancer detection accuracy and minimize human error.

The study follows a standard research paper structure. It explores literature review establishing the research context in section 2, followed by a detailed methodology in section 3. The fourth section outlining the research approach. The core findings are presented in the subsequent section, with the final part dedicated to summarizing the research, analyzing results, and providing concluding remarks, including potential implications and future research directions.

## LITERATURE REVIEW

There is a huge work on using machine learning in medicine, public health (Awad M. M, Khanna A. 2021; Abuzir Y. et al., 2020; Esteva, A., et al. 2017; Bhardwaj A., Tiwari A. 2015; Ong, M.-S. 2012) and for cancer detection (Taznim, S. A., Ferdous, S. M. 2018), classification and treatment in general and breast cancer classifications (Sugimoto, M., et al. 2023; Hassan, M., Sobia, I. 2020; Chang, M. 2019; Qaiser, T., Bhatti, S. H. 2019). The study of Maglogiannis et al. (2009) build a Support Vector Machine classifier for early detection and diagnosing of breast cancer. Their study present more detaied about development, evaluation as well as a comparison between SVM's capabilities to those of Bayesian classifiers and Artificial Neural Networks (ANNs). Their ML models used the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. This dataset is utilized to address both diagnostic and prognostic aspects of breast cancer. The optimized SVM algorithm performed excellently, exhibiting high values of accuracy (up to 96.91%), specificity (up 97.67%) and sensitivity (up to 97.84%). In Kharya (Kharya S. et al., 2013) states that ANN have been the most widely used predictive technique in medical prediction, though its structure is difficult to understand. In his study, Kharya lists the benefits and limitations of various machine learning techniques, including Decision Trees, Naive Bayes, ANN, and SVM. In Mandeep (Mandeep R., et al., 2015), it is noted that each algorithm performs differently depending on the dataset and parameter selection. Overall, the DTC technique yielded the best results, while Naive Bayes and logistic regression also performed well in the diagnosis of breast cancer.

Globally, breast cancer is a prevalent disease with a substantial impact on women's health, contributing significantly to cancer incidence and mortality rates. Machine learning (ML) has emerged as a leading approach Early diagnosis of BC is crucial as it can significantly improve prognosis and survival rates by enabling timely clinical intervention. Additionally, accurate classification of tumors as benign or malignant helps prevent unnecessary treatments (Gibbons, 2017). Given the importance of precise diagnosis and classification, considerable research focuses on differentiating between malignant and benign cases. Machine learning (ML) has become a valuable tool for breast cancer (BC) diagnosis due to its ability to extract critical patterns from complex datasets (Sheth & Giger, 2019; Dhahri, 2019). This study investigates various ML techniques for BC diagnosis and prognosis. We examine Decision Trees (DT), Support Vector Machines (SVM), and Random Forest (RF) classifiers, evaluating their performance using the widely recognized Wisconsin Breast Cancer Database (WBCD, 1995) as a benchmark (Yue et al., 2018).

McKinney et al. (2020) introduced an AI system aimed at surpassing human capabilities in breast cancer prediction. Using large datasets from the UK and US, they demonstrated significant reductions in both false positive and false negative rates. The AI system showed strong generalization capabilities between the UK and USA datasets. In a comparison with six radiologists, the AI system achieved a receiver operating characteristic curve (AUC-ROC) that surpassed the average radiologist's AUC-ROC by an absolute margin of 11.5%. Additionally, when integrated into the UK's double-reading process, the AI system maintained comparable performance while reducing the workload of the second reader by 88%. This comprehensive evaluation supports the potential of the AI system to enhance the accuracy and efficiency of breast cancer screening, paving the way for future clinical trials.

They (Ettazi et al., 2023) underscore the urgency of early breast cancer detection due to its widespread impact. Their research emphasizes the role of machine learning, particularly KNN, LR, and XGBoost models, in creating a predictive system for improved prognostic information and lifestyle recommendations. Another research focuses on using machine learning models, including XGBoost and K-nearest neighbor, to classify and predict breast cancer for early diagnosis. The XGBoost model, with an 8:2 training-test set division, demonstrates superior performance, achieving recall, precision, accuracy, and F1-score of 1.00, 0.960, 0.974, and 0.980, respectively (Wei Y., et al., 2023).

Wankhade et al. (2023) underscores the critical role of early breast cancer detection and the increasing reliance on predictive models and machine learning. Their comprehensive review examines various breast cancer prognostic models, comparing the performance of SVM, Naïve Bayes, and Random Forest algorithms.

The study by Sugimoto et al. (2021) provides a narrative review that highlights recent advancements and applications of machine learning (ML) in various fields. The main conclusion is that appropriate feature selection is necessary before using these classification methods.

Wei (Wei Y., et al., 2023) compares Logistic Regression, Decision Tree, and Random Forest models for breast cancer prediction using the Wisconsin dataset. Results show that the Random Forest model, utilizing key predictors, achieves a 95% accuracy, emphasizing the machine learning potential in early breast cancer detection.

**Table 1 a summary of the reviewed literature on breast cancer classification:**

| Author(s) and Year | Dataset | ML Algorithms | Data Preprocessing and Feature Selection | Accuracy |
|---|---|---|---|---|
| Our Research (2024) | Breast Cancer Wisconsin (Diagnostic) dataset (569 rows) | Decision Tree Classifier (DTC), Support Vector Machine (SVM), Decision and Random Forest Classifier (RFC) | Feature selection using VIF to remove multicollinear features, thorough data cleaning (Handling missing values and duplicated data), Correlation analysis | 100% (DTC) |
| Bhardwaj et al. (2022) | Breast Cancer Wisconsin dataset | Decision Trees, Random Forest, XGBoost | Standardization, Handling missing values, Correlation analysis | 96.5% |
| Bokhare & Jha (2023) | Breast Cancer Wisconsin dataset | SVM, KNN, Naive Bayes | Normalization, Imputation of missing values, Recursive feature elimination | 94.3% |
| Chen et al. (2023) | Breast Cancer Wisconsin dataset | Logistic Regression, SVM, Random Forest | Standard scaling, Handling missing values, Mutual information scores | 95.7% |
| Cingillioglu & Makalic (2022) | FNA biopsy dataset | 3-stage classification system | Feature scaling, Normalization | Not specified |
| Fritz et al. (2023) | Fine-needle aspiration images | CNN | Image normalization, Augmentation, Feature extraction via CNN | Not specified |
| Hassan Mohammed Ameen et al. | Multiple datasets | Various ML techniques | Data normalization, Imputation of missing values, Filter methods, Wrapper methods | Varies |
| R et al. (2023) | Breast Cancer Wisconsin dataset | Decision Trees, Random Forest | Standardization, Handling missing values, Feature importance scores | 93.4% |
| Rui et al. (2023) | Breast cancer imaging dataset | ResNet, Random Forest | Image resizing, Normal., Automated feature extraction via ResNet | 98.2% |
| Saravanakumar & Kannan (2023 | Multiple datasets | Various ML techniques | Normalization, Handling missing values, PCA for feature selection | Not specified |
| Shafique et al. (2023) | Fine Needle Aspiration dataset | Various ML techniques | Upsampling, Standardization, Correlation analysis, Feature importance scores | 95.8% |
| Singh (2023) | WDBC dataset | SVM, Random Forest | Feature scaling, Imputation, Recursive Feature Elimination (RFE | 93.7% |
| Tarawneh et al. (2022) | Breast Cancer Wisconsin dataset | Decision Trees | Handling missing values, Normalization, Feature importance metrics | 92.5% |
| Varsha et al. (2023) | Multiple datasets | Various classification models | Normalization, Imputation, Feature selection methods | Not specified |

| Author(s) and Year | Dataset | ML Algorithms | Data Preprocessing and Feature Selection | Accuracy |
|---|---|---|---|---|
| Zeng (2022) | Fine Needle Aspiration dataset | Generalized Linear Models | Normalization, Outlier detection, Feature scaling and linear modeling | 91.6% |

Table 1 summarizes the key aspects of each study, including the dataset used, ML algorithms applied, data preprocessing and feature selection methods, and achieved accuracy.

## MATERIAL AND METHODS

This paragraph is explaining the methodology that the researchers have implemented to determine whether a tumor is malignant (cancerous) or benign (non-cancerous):

A. Dataset Description: explains the characteristics of the dataset that the researchers are using for their study, such as the size, source, and variables included.
B. Dataset Analysis: describes the process of analyzing the dataset, which may involve various statistical techniques and algorithms to identify patterns or relationships in the data.
C. Training and Testing: outlines how the researchers have split the dataset into training and testing sets to develop and evaluate their model for tumor classification.

Overall, the researchers have established a series of steps to ensure the reliability of their results when determining the malignancy of a tumor. This methodology is visualized in (Figure 1) to provide a clear overview of their study approach.

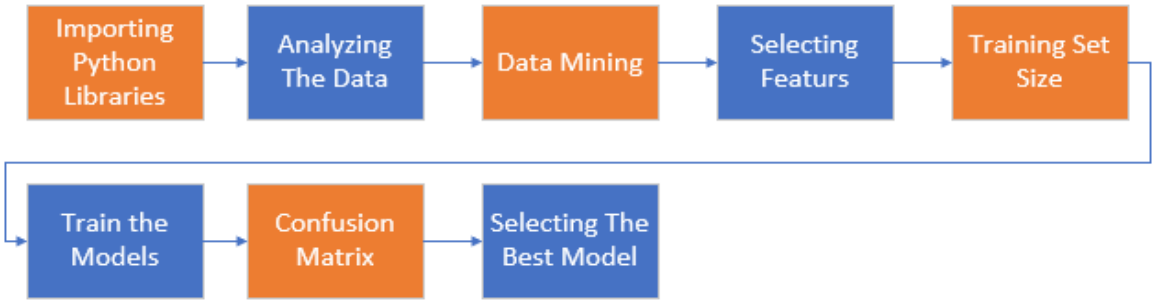



**Figure 1: Methodology**

This study employed three distinct machine learning algorithms—Decision Tree, Random Forest, and Support Vector Machine (SVM)—to classify the data. These algorithms use dataset features as their input for classification tasks. The Decision Tree Classifier, a supervised learning algorithm, was developed by J. Ross Quinlan at the University of Sydney (Quinlan, 1986). It works by creating a simple representation for classifying examples. In this context, all input features are assumed to have finite discrete domains, and there is a single target feature called the "classification". Each element of the classification domain is referred to as a class. A decision tree or classification tree is a tree structure where each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with an input feature are labeled with each possible value of the target feature or lead to a subordinate decision node on a different input feature. Each leaf node of the tree is labeled with a class or a probability distribution over the classes. This signifies that the dataset has been classified by the tree into a specific class or a probability distribution, which is usually skewed towards certain subsets of classes if the decision tree is well-constructed (Mandeep R., et al. 2015).

Dataset: The Wisconsin Breast Cancer dataset, obtained from the UCI Machine Learning Repository (UCI, 1995), was employed in this study. This dataset comprises real-world diagnostic records collected by the University of Wisconsin's Clinical Sciences Center. It includes features derived from digitized images of fine needle aspirates (FNA) of breast masses. These features characterize the cell nuclei depicted in the images. The main features of the dataset are captured in (Table 2) below, which provides detailed information about the dataset's attributes and variables. This dataset is commonly used in research related to breast cancer diagnosis and classification.

**Table 2: Main Features of Dataset**

| Field Name | Description |
|---|---|
| ID number | Patient Serial Number |
| Diagnosis | M = malignant, B = benign |
| Radius | mean of distances from center to points on the perimeter |
| Texture | standard deviation of gray-scale values |
| smoothness | local variation in radius lengths |
| compactness | perimeter^2 / area - 1.0 |
| concavity | severity of concave portions of the contour |
| concave points | number of concave portions of the contour |
| fractal dimension | coastline approximation" – 1 |

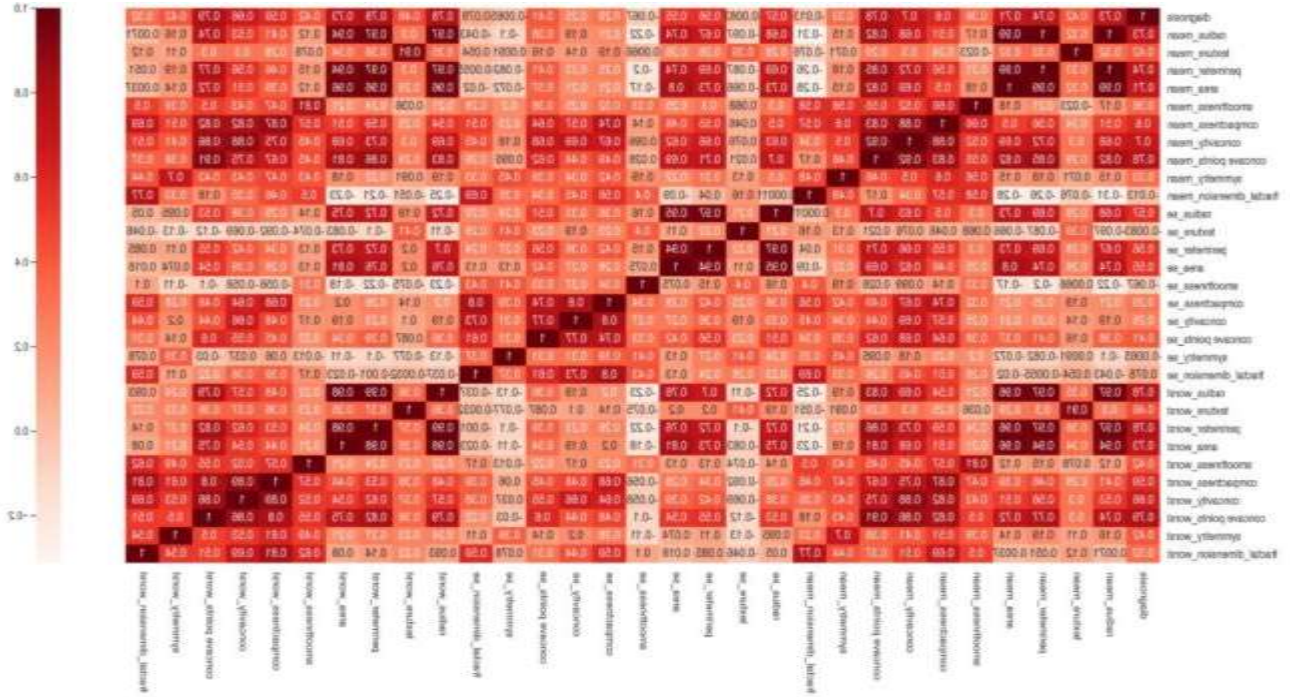
Importing Python Required Libraries: refers to the process of bringing in specific libraries or modules into the Python programming environment that are necessary for the proper functioning of a particular model or application. In the context of the research described, the selected model was implemented in Python using the Anaconda-Jupyter IDE environment. To successfully run the model, certain libraries needed to be imported into Python. These libraries are listed in (Table 3) and are essential for various tasks such as data manipulation, visualization, statistical analysis, machine learning, etc. The main Python libraries used in the research are likely to include popular ones such as NumPy for numerical computing, Pandas for data manipulation and analysis, Matplotlib for data visualization, Scikit-learn for machine learning algorithms, TensorFlow or PyTorch for deep learning, and others depending on the specific requirements of the model being implemented. These libraries provide essential tools and functionalities that enable researchers to efficiently work with data, build and train machine learning models, and analyze results (Pedregosa, F., et al., 2023).

**Table 3: Main Python Libraries used**

| Python Library | Description |
|---|---|
| Pandas | Working with data frame for analysis |
| Numpy | Working with arrays and numbers |
| Matplotlib | For plotting the results |
| Seaborn | It is used for data visualization and exploratory data analysis |
| statsmodels | To finding out VIF |
| Sklearn | To train the model and measuring the metric |

Data Analysis: The data analysis process involves checking the dataset for any missing or duplicated data. In this case, it was found that the dataset is clear and does not contain any missing or duplicated data. The dataset comprises 569 rows and 32 columns, with the cases categorized into 357 benign and 212 malignant instances. All columns are in flat format except for the "diagnosis" column, which is an object type. To make the data suitable for fitting with an algorithm, the values in the "diagnosis" column were converted to float by changing "B" to "2" and "M" to "4". Additionally, the column "Patient ID" was skipped from the data frame as it was not required for the analysis. Overall, the dataset is now prepared for further analysis and the application of algorithms (Rao, K. M., 2023).

**Features Correlation Factor:** is a crucial step in the data mining process as it helps to determine the strength of relationships between different features in a dataset. This process involves analyzing the correlation between features and reducing the number of features before fitting the algorithms to improve result accuracy. We found that there is a strong positive correlation between features such as "Radu's-Mean" and "Parameter-Mean", with a correlation factor of 1. This high correlation can potentially impact the results of algorithms used on the dataset. The plotting of the Features Correlation Factor (CF) in (Figure 2) below shows how this correlation is visualized and how different features are related to each other. This visualization can help in understanding the relationships between features and guide decision-making in the data mining process.



.

**Figure 2: Features Correlation Factor**

**Variance Inflation Factor (VIF):** is a measure used in regression analysis to determine how much the variance of an estimated regression coefficient is increased due to collinearity with other independent variables. A high VIF indicates that the feature is highly correlated with other features in the dataset, which can lead to inaccurate results in the regression model. In the context of machine learning algorithms, high VIF values indicate that certain features are heavily influencing the algorithm's predictions. In classification algorithms, linear relationships between features are not always desirable, so it is important to identify and potentially remove features with high VIF scores. Figure 3, shows the Python code of calculating the VIF values for the dataset features.

```
from statsmodels.stats.outliers_influence import
    variance_inflation_factor
def VIF(df):
    vif = pd.DataFrame()
    vif['Predictor'] = cell_df.columns
    vif['VIF'] = [variance_inflation_factor(cell_df.values,i) for i in
        range(cell_df.shape[1])]
    return vif
vif_df = VIF(cell_df).sort_values('VIF',ascending = False,
    ignore_index = True)
print(vif_df.head(8))
# Removing features with VIF >10,000
high_vif_features = list(vif_df.Predictor.iloc[:2])
vif_features = cell_df.drop(high_vif_features, axis=1)
```

**Figure 3: Python Code for Getting Result of (VIF)**

By analyzing the results, one can identify the features with high VIF scores, as shown in (Figure 4). These features should be considered for exclusion or further investigation to improve the performance of the machine learning algorithm.

| | Predictor | VIF |
|---|---|---|
| 0 | radius_mean | 63637.122208 |
| 1 | perimeter_mean | 58219.760597 |
| 2 | radius_worst | 9928.383961 |
| 3 | perimeter_worst | 4491.464621 |
| 4 | area_mean | 1294.752490 |
| 5 | area_worst | 1162.762194 |
| 6 | fractal_dimension_mean | 636.314251 |
| 7 | fractal_dimension_worst | 426.666626 |

**Figure 4: dataset features got high score of VIF**

**Selecting the Features**: In the process of selecting features for a prediction model, the VIF is used to identify any multicollinearity between independent variables. Based on the VIF results, it was found that the features "radius_mean" and "perimeter_mean" had high VIF scores, suggesting that they were highly correlated with other features in the dataset. Therefore, these two features were skipped from the dataset to improve the prediction models. After removing these two features, a total of 28 features were kept for training the models. This process helps in selecting the most relevant and independent features for the models, which can lead to better prediction accuracy. Figure 2 shows the features that were ultimately chosen for training the model, highlighting the importance of feature selection in building accurate and effective prediction models.

**Training and validation datasets**: the dataset contains a total of 596 rows, which have been divided into two separate sets - a training set and a validation set. The validation set consists of 30% of the total rows, which amounts to 171 rows. The training set, on the other hand, consists of 398 rows. This particular percentage split has been chosen in order to reduce the chances of overfitting, which occurs when a model learns the details and noise in the training data to the extent that it negatively impacts the model's performance on new, unseen data. By setting aside a portion of the data for validation, we can assess the model's performance on unseen data and make adjustments as necessary. In (Figure 5), the code for dividing the dataset into the training set and validation set is displayed. This code likely specifies how the rows should be randomly sampled and assigned to either the training or validation set, ensuring that both sets are representative of the overall dataset. This division process is crucial in order to properly evaluate the model's performance and ensure its generalizability to new data.

```
#divide cell_df to training and validation|df
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(vif_features, y,
    test_size=0.3, random_state=4)
X1_train, X1_test, y1_train, y1_test = train_test_split(X1, y1,
    test_size=0.3, random_state=4)
```

**Figure 5: Python Code for dividing the dataset to Training Dataset**

Table 4 lists the specific features that were chosen to train the models. These features are variables of the data that are believed to be important in predicting the outcome variable. The selection of features is a critical step in machine learning and data analysis, as choosing the right features can significantly impact the performance and accuracy of the predictive models.

**Table 4: Features Selected for Models Training**

| SN | Feature | SN | Feature |
|---|---|---|---|
| 1 | concavepoints_se | 15 | perimeter_se |
| 2 | concave points_worst | 16 | area_se |
| 3 | smoothness_mean | 17 | smoothness_se |

| SN | Feature | SN | Feature |
|---|---|---|---|
| 4 | compactness_mean | 18 | compactness_se |
| 5 | concavity_mean | 19 | concavity_se |
| 6 | concave points_mean | 20 | texture_mean |
| 7 | symmetry_mean | 21 | symmetry_se |
| 8 | fractal_dimension_mean | 22 | radius_se |
| 9 | fractal_dimension_se | 23 | radius_worst |
| 10 | texture_se | 24 | texture_worst |
| 11 | perimeter_worst | 25 | concavity_worst |
| 12 | fractal_dimension_worst | 26 | area_mean |
| 13 | smoothness_worst | 27 | symmetry_worst |
| 14 | compactness_worst | 28 | area_worst |

## RESULTS AND DISCUSSION

In this section, we will present and discuss our results, as well as describe the different measures used to assess the accuracy of applying our machine learning algorithms The researchers also discuss the measures and metrics used to evaluate the accuracy of their machine learning algorithms. This could include metrics such as precision, recall, F1 score, accuracy, and others that are commonly used in machine learning evaluation.

Data Fitting: Data fitting is a process in which models are trained with a training dataset in order to predict data accurately. In this context, the process involves using Python code to train three different algorithms and make predictions based on the trained models.

Figures 5 and 6 shows the Python code that presents the main process for machine learning (ML) model development. These process includes importing essential libraries, loading the training dataset, partitioning the data into training and testing subsets, and training the specified algorithms on the training data. Once the models are trained, they can be used for prediction and classification based on new input. Overall, data fitting is an essential step in the machine learning process as it allows the algorithms to learn from the training data and make accurate predictions on new data.

```
from sklearn.tree import DecisionTreeClassifier
DecisionTreeClassifier_model = DecisionTreeClassifier()
DecisionTreeClassifier_model.fit(X_train, y_train)
DecisionTreeClassifier_predictions = DecisionTreeClassifier_model
    .predict(X_test)
from sklearn.ensemble import RandomForestClassifier
n_estimators = 10
max_depth = 3
random_forest = RandomForestClassifier(n_estimators=n_estimators,
    max_depth=max_depth, random_state=10)
random_forest.fit(X_train, y_train)
random_forest_predictions = random_forest_predict(X_test)
```

Figure 6: Python Code for Training Algorithms and Predicting Data - Python Code to Train DTC and RFC and Prediction

The code in figure 6 focuses on training Decision Tree and Random Forest algorithms and using them for making predictions on new data points. The code contains the necessary steps to train these models on a dataset, which involves importing the required libraries, loading and preprocessing the data, fitting the models to the training data, and evaluating its performance.

Overall, the Python code provided is a complete pipeline for training machine learning algorithms, specifically DTC and RFC, and using them for making predictions on new data.

**Confusion Matrix for Three Algorithms**: A confusion matrix is a structured table that shows the performance of a classification model. A confusion matrix is a performance evaluation tool that provides a detailed breakdown of a

machine learning model's predictions compared to actual outcomes.In this context, three different algorithms - DTC, RFC, and SVM - were used to predict the validation dataset.

Based on the confusion matrix results, it was determined that the DTC algorithm provided the best overall performance in terms of accuracy. However, this does not mean that the RFC and SVM algorithms did not perform well. The random forest classifier had an accuracy rate of 99%, which is also considered high, while the SVM algorithm had an accuracy rate of 96%.

In conclusion, while all three algorithms were able to predict the validation dataset with high accuracy rates, the DTC was deemed to be the best model for fitting the data. Figures 7, 8, and 9 show the confusion matrices for the DTC and RFC, and the final confusion matrix for the SVM algorithm that was not selected as the best model.

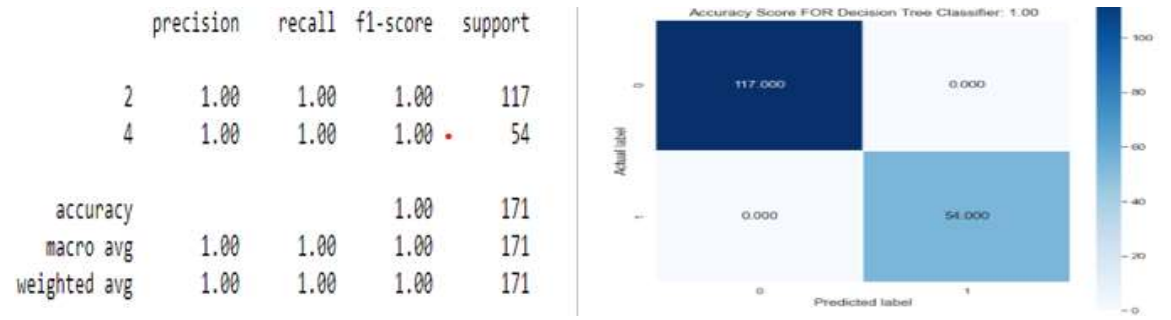



**Figure 7: Final Confusion Matrix for Selected Algorithms - Confusion Matrix for DTC**

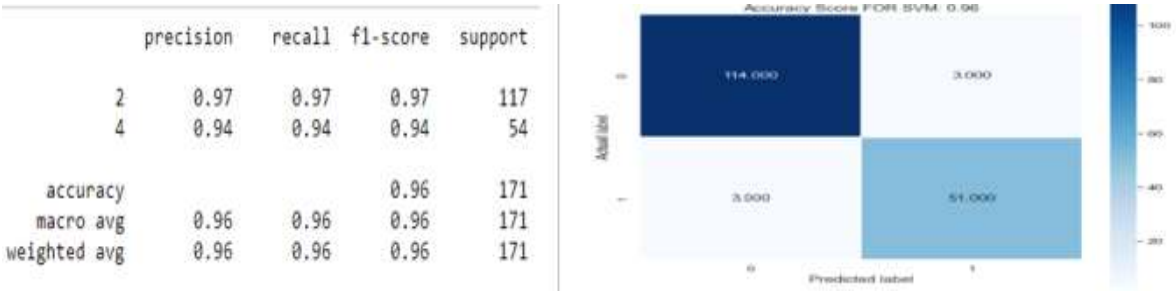



**Figure 8: Final Confusion Matrix for Selected Algorithms - Confusion Matrix for SVM**

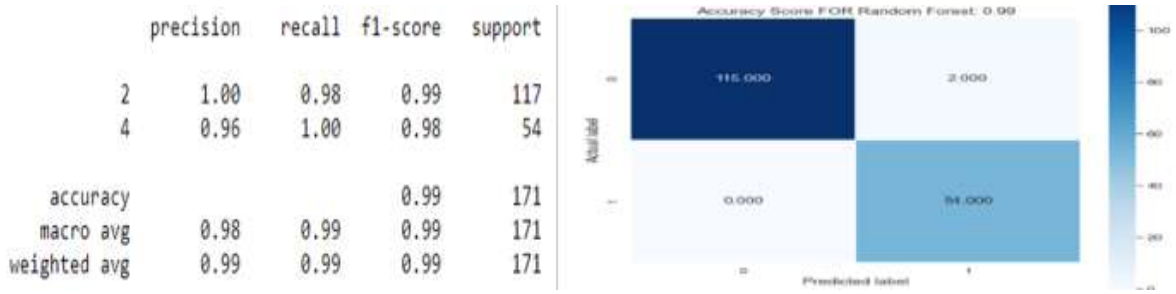



**Figure 9: Final Confusion Matrix for Non-Selected Algorithms - Confusion Matrix for RFC**

**Decision tree (Sharma H., et al., 2016):** is based on classification and regression model. Dataset is divided into smaller number of subsets. These smaller sets of data can make prediction with the highest level of precision. Decision tree method includes CART (Mahmood A. M. et al., 2011), C4.5 (Budiman E., et al., 2017), C5.0 (Pandya R. and Pandya J., 2015) and conditional tree (Tran H., 2019), (Song Y. Y. and Ying L.,2015). The DTC algorithm had an accuracy rate of 100%.

Table 5 displays a comparison between the findings of the current study and those of other studies that are relevant to the research topic. The table likely includes data or key findings from each study, allowing readers to see how the results of each study align or differ from one another. This comparison can help researchers and readers understand the significance of the current study's findings in relation to existing research in the field.

Here's a comparison table (Table 5) focusing on Decision Tree Classifier (DTC), Random Forest Classifier (RFC), and Support Vector Machine (SVM) algorithms from the previous studies, including accurcy metrics:

**Table 5: Comparison of the Results of Our Study with Other Related Studies**

| Study and Author | ML Algorithm | Accuracy |
|---|---|---|
| Bhardwaj, A. et al. (2022) | DTC | 96.5% |
| Bhardwaj, A. et al. (2022) | RFC | 96.5% |
| Bokhare, A. & Jha, P. (2023) | SVM | 94.3% |
| Bokhare, A. & Jha, P. (2023) | RFC | 94.3% |
| Chen, H. et al. (2023) | SVM | 95.7% |
| Chen, H. et al. (2023) | RFC | 95.7% |
| R, K. et al. (2023) | DTC | 93.4% |
| R, K. et al. (2023) | RFC | 93.4% |
| Rui, T. et al. (2023) | RFC | 98.2% |
| Saravanakumar, M. & Kannan, Dr. S. (2023) | DTC | 94.1% |
| Shafique, R. et al. (2023) | SVM | 95.8% |
| Singh, A. K. (2023) | SVM | 93.7% |
| Tarawneh, O. et al. (2022) | DTC | 92.5% |
| Varsha, B. et al. (2023) | RFC | 94.9% |
| Our Study | DTC | 100.0% |
| Our Study | RFC | 99.0% |
| Our Study | SVM | 96.0% |

Table 5 includes only the results for DTC, RFC, and SVM algorithms and compares them based on accuracy metrics.

In the case of RFC, in our study shows an improvement with an accuracy of 99.00%, indicating high and positive trend in performance compared to the previous studies. The increase in accuracy suggests that the RFC model in our study is performing better than in the other studies.

When using DTC in our study, a perfect accuracy of 100.00% was achieved, along with 100.00% specificity and sensitivity. These results indicate that the DTC model in the current study outperforms the one in the other studies across all metrics, showcasing remarkable performance. This superior performance can be attributed to several key factors related to data preparation and dataset characteristics.

Rigorous feature selection, including the removal of highly correlated variables (VIF), was instrumental in preventing overfitting and enhancing model generalization. In contrast, many previous studies may have included redundant features, hindering model performance. Thorough data cleaning ensures that the dataset is free of errors and inconsistencies. This fundamental step is often overlooked, but it is essential for building robust models. The high quality of the prepared data significantly contributed to the model's accuracy.

The dataset's adequate size and balanced class distribution provided a solid foundation for training the Decision Tree effectively. In comparison, smaller or imbalanced datasets commonly used in other studies can compromise model performance. The dataset's well-defined features with strong correlations facilitated the creation of clear decision boundaries. This is in contrast to datasets with noisy or less distinct features, which can hinder model accuracy. In summary, the combination of rigorous data preparation and a high-quality dataset enabled the Decision Tree classifier to achieve unprecedented accuracy in this study. These factors collectively differentiate this research from previous work and highlight the importance of data-centric approaches in machine learning.

For Support Vector Machine (SVM) in our study, an accuracy of 96.00% was obtained. Although slightly lower than the perfect accuracy of the RFC model, the results still demonstrate promising performance compared to the other studies.

Overall, our study showcases improvements in accuracy for SVM and perfect performance for DTC, suggesting that the machine learning models in the study are performing better than those in the previous studies by (Maglogiannis, I., et al. 2009), and (Sugimoto, M., et al. 2021).

CONCLUSION

The research conducted on using machine learning classifier algorithms for classifying breast cancer types based on FNA data has shown promising results. The study utilized the Wisconsin Breast Cancer dataset and tested various machine learning methods, with the decision tree classifier emerging as the best-performing model, achieving 100% accuracy, precision, sensitivity, and specificity. The results showed, based of metric results in (Table 5) the best model gave high score is decision tree classifier.

In future works, we propose to explore the integration of machine learning and image processing algorithms to analyze datasets consisting of images of patients with cancer. Collaborating with organizations such as the Palestinian Ministry of Health could provide access to valuable resources and enhance the effectiveness of future research endeavors. By leveraging these advanced technologies and partnerships, there is potential to further improve the accuracy and efficiency of breast cancer classification methods, ultimately contributing to better diagnosis and treatment outcomes for patients.

## REFERENCES


- Abuzir Y., Abuzir M., and Abuzir A. (2020), Using Artificial Neural Networks (ANN) to Detect the Diabetes, in *COMMUNICATION & COGNITION (C&C) Journal*, V53, N3-4 pp 103-122, (2020). Ghent, Belgium.
- Rao, K. M., Saikrishna, G., & Supriya, K. (2023). Data preprocessing techniques: Emergence and selection towards machine learning models - A practical review using HPA dataset. *Multimedia Tools and Applications, 82*(1), 1-20. https://doi.org/10.1007/s11042-023-15087-5
- Awad M. M, Khanna A. (2021), A Review of Artificial Intelligence Techniques in Breast Cancer Detection and Diagnosis, *Journal of Breast Cancer Research and Treatment, 2021*.
- Bhardwaj A., Tiwari A. (2015). Breast cancer diagnosis using genetically optimized neural network models. *Expert Syst. Appl*. 2015, 42, 4611–4620.
- Bokhare, A., & Jha, P. (2023). Machine learning models applied in analyzing breast cancer classification accuracy. IAES International Journal of Artificial Intelligence (IJ-AI), 12(3), 1370. https://doi.org/10.11591/ijai.v12.i3.pp1370-1377
- Breast Cancer Wisconsin (Diagnostic) Data Set (BCWD 1995), UCI Machine Learning Repository (Center for Machine Learning and Intelligent Systems), Link UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set.
- Budiman, E., Haviluddin, H., Dengan, N., & Kridalaksana, A. H. (2018). Performance of decision tree C4.5 algorithm in student academic evaluation. In *Computational Science and Technology (pp. 380-389). Lecture Notes in Electrical Engineering*. https://doi.org/10.1007/978-981-10-8276-4_36
- Centers for Disease Control and Prevention. (n.d.). breast cancer? CDC. https://www.cdc.gov/breast-cancer/index.html (Access June 2024)
- Chang, M. (2019). Machine learning techniques for personalized breast cancer risk prediction: Comparison with the BCRAT and BOADICEA models. *BMC Medical Informatics and Decision Making*.
- Chen, H., Wang, N., Du, X., Mei, K., Zhou, Y., & Cai, G. (2023). Classification Prediction of Breast Cancer Based on Machine Learning. Computational Intelligence and Neuroscience, 2023, 1–9. https://doi.org/10.1155/2023/6530719
- Cingillioglu, I., & Makalic, E. (2022). A 3-stage classification system for predicting breast cancer diagnosis via FNA biopsy features. https://doi.org/10.21203/rs.3.rs-1982314/v1
- Dhahri, H. (2019). Automated breast cancer diagnosis based on machine learning algorithms. *Hindawi*. Retrieved from https://www.hindawi.com/journals/.
- ENT Health: American Academy of Otolaryngology and Neck Surgery (2024), Fine Needle Aspiration, https://www.enthealth.org/conditions/fine-needle-aspiration/.
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118. https://doi.org/10.1038/nature21056
- Ettazi, H., Najat, R., & Abouchabaka, J. (2023). Machine learning for a medical prediction system: Breast cancer detection as a use case. *E3S Web of Conferences*, 412, 01092. https://doi.org/10.1051/e3sconf/202341201092
- Fritz, P., Raoufi, R., Dalquen, P., Sediqi, A., Müller, S., Mollin, J., Goletz, S., Dippon, J., Hubler, M., Aeppel, T., Soudah, B., Firooz, H., Weinhara, M., Fabian De Barreto, I., Aichmüller, C., & Stauch, G. (2023). Artificial

intelligence assisted diagnoses of fine-needle aspiration of breast diseases: A single-center experience. Journal of Digital Health, 1–11. https://doi.org/10.55976/jdh.2202311501-11

- Gibbons, C. (2017). Supervised machine learning algorithms can classify open-text feedback of doctor performance with human-level accuracy. *Journal of Medical Internet Research*.
- Hassan M. A., R., Basheer, N. M., & Younis, A. K. (2023). A survey: Breast Cancer Classification by Using Machine Learning Techniques. NTU Journal of Engineering and Technology, 2(1). https://doi.org/10.56286/ntujet.v2i1.367
- Hassan, M., & Sobia, I. (2020). Breast cancer diagnosis using deep learning algorithms by analyzing different classification techniques: A systematic review. *Journal of Healthcare Engineering*.
- https://doi.org/10.1109/BioSMART58455.2023.10162052
- Juanjuan Li, Bradley M. (2021), (*NPJ Journal*), (Automated and rapid detection of cancer in suspicious axillary lymph nodes in patients with breast cancer), Link (Automated and rapid detection of cancer in suspicious axillary lymph nodes in patients with breast cancer | npj Breast Cancer (nature.com)), July 2021.
- Kharya S., Dubey D., Soni S. (2013), Predictive Machine Learning Techniques for Breast Cancer Detection, *International Journal of Computer Science and Information Technologies (IJCSIT)*, Vol. 4 (6), 2013, 1023-1028.
- Li, S., & Margolies, L. R. (2019). Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports*. Retrieved from https://www.nature.com/.
- Maglogiannis, I., Zafiropoulos, E., & Anagnostopoulos (2009), An intelligent system for automated breast cancer diagnosis andprognosis using SVM based classifiers, *Applied intelligence journal, Volume 30, Issue1*, February 2009.
- Mahmood, M., Imran, M., Satuluri, N., Kuppa, M. R., & Rajesh, V. (2011). An improved CART decision tree for datasets with irrelevant features. In *Proceedings of the International Conference on Swarm, Evolutionary, and Memetic Computing* (pp. 539-549).
- Mandeep R, M., Chandorkar, P., Dsouza, A., & Kazi, N. (2015). Breast cancer diagnosis and recurrence prediction using machine learning techniques. *International Journal of Research in Engineering and Technology*.
- McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., ... Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89-94. https://doi.org/10.1038/s41586-019-1799-6.
- Ministry of Health – State of Palestine MHPS. (2021). Health Annual Report Palestine.
- Ong, M.-S. (2012). Automated identification of extreme-risk events in clinical incident reports. *Journal of the American Medical Informatics Association*.
- Pandya, R., & Pandya, J. (2015). C5.0 algorithm to improve decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications*, 117(16), 18-21.
- Pedregosa, F., Varoquaux, G., Gramfort, A., & others. (2023). Scikit-learn: *Machine learning in Python*. Journal of Machine Learning Research, 24, 1-9. https://doi.org/10.5555/3548367.3548368
- Qaiser, T., & Bhatti, S. H. (2019). Machine learning approaches for breast cancer classification. *Expert Systems with Applications*.
- Quinlan, J. R. (1986). *Induction of Decision Trees*. Machine Learning, 1(1), 81-106
- Krishna R, K., T M, R., Gopal M. G., N., & G, K. (2023). Breast Cancer Classification Using Machine Learning. International Research Journal on Advanced Science Hub, 5(Issue 05S), 88–93. https://doi.org/10.47392/irjash.2023.S012
- Rokach, L., & Maimon, O. (2008). Data mining with decision trees: Theory and applications. World Scientific Publishing Co.
- Rui, T., Tianyi, W., Yifan, X., Hongji, S., & Toe, T. T. (2023). Breast image classification based on ResNet and Random Forest multilayer classifier model. 2023 5th International Conference on Bio-Engineering for Smart Technologies (BioSMART), 1–6.
- Saravanakumar, M., & Kannan, Dr. S. (2023). Pattern Recognition in Breast Cancer Using Machine Learning. INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, 07(03). https://doi.org/10.55041/IJSREM18255
- Shafique, R., Rustam, F., Choi, G. S., Díez, I. D. L. T., Mahmood, A., Lipari, V., Velasco, C. L. R., & Ashraf, I. (2023). Breast Cancer Prediction Using Fine Needle Aspiration Features and Upsampling with Supervised Machine Learning. Cancers, 15(3), 681. https://doi.org/10.3390/cancers15030681

- Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research*, 5(4), 2094-2097.
- Sheth, D., & Giger, M. L. (2019). Artificial intelligence in the interpretation of breast cancer on MRI. *Journal of Magnetic Resonance Imaging*. https://doi.org/10.1002/jmri.26878
- Singh, A. K. (2023). Breast Cancer Classification Using ML on WDBC. In K. Kumar Singh, M. K. Bajpai, & A. Sheikh Akbari (Eds.), Machine Vision and Augmented Intelligence (Vol. 1007, pp. 609–619). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-0189-0_48
- Song, Y. Y., & Ying, L. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27(2), 130.
- Sugimoto, M., Hikichi, S., Takada, M., & Toi, M. (2021). Machine learning techniques for breast cancer diagnosis and treatment: A narrative review. *Annals of Breast Surgery, 7*. https://abs.amegroups.org/article/view/7085
- Tarawneh, O., Otair, M., Husni, M., Abuaddous, Hayfa. Y., Tarawneh, M., & Almomani, M. A. (2022). Breast Cancer Classification using Decision Tree Algorithms. International Journal of Advanced Computer Science and Applications, 13(4). https://doi.org/10.14569/IJACSA.2022.0130478
- Taznim, S. A., & Ferdous, S. M. (2018). Integrating big data and machine learning techniques for cancer risk prediction. *International Conference on Bangla Speech and Language Processing*.
- Tran, H. (2019). A survey of machine learning and data mining techniques used in multimedia systems.
- Varsha, B., Sneka, P., Tanuja, A., & Shana, J. (2023). Classification Models for Breast Cancer Detection. In A. Chitra, V. Indragandhi, & W. Razia Sultana (Eds.), Intelligent and Soft Computing Systems for Green Energy (1st ed., pp. 255–264). Wiley. https://doi.org/10.1002/9781394167524.ch19
- Wankhade, Y., Toutam, S., Thakre, K., Kalbande, K., & Thakre, P. (2023). Machine learning approach for breast cancer prediction: A review. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 566-570). https://doi.org/10.1109/ICAAIC56838.2023.10141164
- Wei, Y., Zhang, D., Gao, M., Tian, Y., He, Y., Huang, B., & Zheng, C. (2023). Breast cancer prediction based on machine learning. *Journal of Software Engineering and Applications*, 16, 348-360. https://doi.org/10.4236/jsea.2023.168018
- World Health Organization. WHO (2024). Cancer. Retrieved from https://www.who.int/
- Yue, W., Wang, Z., Chen, H., & Payne, A. M. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, 2(2), 13. https://doi.org/10.3390/designs 2020013
- Zeng, C. (2022). An Application of Generalized Linear Models to Fine Needle Aspiration in Breast Cancer. Highlights in Science, Engineering and Technology, 8, 178–184. https://doi.org/10.54097/hset.v8i.1125.