Big Data processing methods for environmental management.

Bc. George Amoako Danquah

Master's thesis

2023

Tomas Bata University in Zlín
Faculty of Applied Informatics

Tomas Bata University in Zlín
Faculty of Applied Informatics
Department of Informatics and Artificial Intelligence

Academic year: 2022/2023

ASSIGNMENT OF DIPLOMA THESIS

(project, art work, art performance)

Name and surname:	**George Amoako Danquah**
Personal number:	**A19886**
Study programme:	**N3902 Engineering Informatics**
Branch:	**Information Technologies**
Type of Study:	**Full-time**
Work topic:	**Metody zpracování velkých dat pro environmentální management**
Work topic in English:	**Big Data Processing Methods for Environmental Management**

Theses guidelines

1. Analyze the literary and professional sources of the topic.
2. Define the main goal of the work.
3. Design an engineering way to solve the assignment.
4. Implement your solution in accordance with the defined goal.
5. Evaluate the success of your project.

Form processing of diploma thesis: **printed/electronic**

Recommended resources:

1. BLANCHARD, Benjamin S. and John BLYLER. System engineering management. Fifth edition. Hoboken, New Jersey: Wiley, 2016. ISBN 9781119047827.
2. Managementmania.com: Professional social network for business [online], Copyright © 2011 – 2016. Wilmington, New Castle County Delaware, USA: MANAGEMENTMANIA.COM [cit. 2022-01-13]. Available from: https://managementmania.com
3. Software Engineering – Ian Sommerville: Tenth Edition [online], 2015. [cit. 2022-01-13]. Available from: https://iansommerville.com/software-engineering-book/

Supervisors of diploma thesis: **prof. Mgr. Roman Jašek, Ph.D., DBA**
Department of Informatics and Artificial Intelligence

Date of assignment of diploma thesis: **December 2, 2022**
Submission deadline of diploma thesis: **May 26, 2023**

Tomas Bata University in Zlín
Faculty of Applied Informatics
L.S.

doc. Ing. Jiří Vojtěšek, Ph.D. m.p.
Dean

prof. Mgr. Roman Jašek, Ph.D., DBA m.p.
Head of Department

In Zlín December 7, 2022

I hereby declare that:

- I understand that by submitting my Diploma thesis, I agree to the publication of my work according to Law No. 111/1998, Coll., On Universities and on changes and amendments to other acts (e.g. the Universities Act), as amended by subsequent legislation, without regard to the results of the defence of the thesis.
- I understand that my Diploma Thesis will be stored electronically in the university information system and be made available for on-site inspection, and that a copy of the Diploma/Thesis will be stored in the Reference Library of the Faculty of Applied Informatics, Tomas Bata University in Zlin, and that a copy shall be deposited with my Supervisor.
- I am aware of the fact that my Diploma Thesis is fully covered by Act No. 121/2000 Coll. On Copyright, and Rights Related to Copyright, as amended by some other laws (e.g. the Copyright Act), as amended by subsequent legislation; and especially, by §35, Para. 3.
- I understand that, according to §60, Para. 1 of the Copyright Act, TBU in Zlin has the right to conclude licensing agreements relating to the use of scholastic work within the full extent of §12, Para. 4, of the Copyright Act.
- I understand that, according to §60, Para. 2, and Para. 3, of the Copyright Act, I may use my work - Diploma Thesis, or grant a license for its use, only if permitted by the licensing agreement concluded between myself and Tomas Bata University in Zlin with a view to the fact that Tomas Bata University in Zlín must be compensated for any reasonable contribution to covering such expenses/costs as invested by them in the creation of the thesis (up until the full actual amount) shall also be a subject of this licensing agreement.
- I understand that, should the elaboration of the Diploma Thesis include the use of software provided by Tomas Bata University in Zlin or other such entities strictly for study and research purposes (i.e. only for non-commercial use), the results of my Diploma Thesis cannot be used for commercial purposes.
- I understand that, if the output of my Diploma Thesis is any software product(s), this/these shall equally be considered as part of the thesis, as well as any source codes, or files from which the project is composed. Not submitting any part of this/these component(s) may be a reason for the non-defence of my thesis.

I herewith declare that:

- I have worked on my thesis alone and duly cited any literature I have used. In the case of the publication of the results of my thesis, I shall be listed as co-author.
- That the submitted version of the thesis and its electronic version uploaded to IS/STAG are both identical.

In Zlin; dated: ………………………………. ....

Student´s Signature

ABSTRAKT

Prognózy tvorby pevných odpadů lze považovat za jednu z největších výzev spojených s environmentálním managementem, zejména pro rozvojové země, kde je sběr dat omezený. Cílem této studie bylo použít umělé neuronové sítě, algoritmus strojového učení, který je metodou zpracování velkých dat, k vytvoření modelu předpovědi produkce odpadu na pevném odpadu v Ghaně na základě dat ze socioekonomických a demografických faktorů. Zpracování a integrace dat byla vyvinuta v softwaru MATLAB. Pro přístup k výkonnosti modelů byly použity ukazatele hodnocení výkonnosti, jako je regrese (R) a střední kvadratická chyba (MSE). Výsledky ukázaly, že umělé neuronové sítě lze použít k vytvoření modelů predikce odpadu a lze je považovat za efektivní přístup k odhadu množství produkovaného odpadu. Očekává se, že výsledky této studie budou představovat obecný přehled pro zainteresované strany v oblasti environmentálního managementu v Ghaně a tyto výsledky lze rozšířit na podobné rozvojové země po celém světě.

Klíčová slova: Umělé neuronové sítě, tvorba odpadů, environmentální management.

ABSTRACT

Solid waste generation forecasting can be considered as one of the biggest challenges associated with environmental management, especially for developing countries where data collection is limited. The aim of this study was to use Artificial Neural Networks, a machine learning algorithm which is a Big Data processing method to create a waste generation forecasting model on Solid waste in Ghana based on data from socio-economic and demographic factors. The processing and integration of data was developed in MATLAB software. Performance assessment indicators such as Regression (R) and Mean Square Error (MSE) were used to access the performance of the models. The results showed that Artificial Neural Networks can be used to create waste prediction models and can be considered as an effective approach to estimating waste generation quantities. The results of this study are expected to represent a general outline for Environmental management stakeholders in Ghana and these results can be extended to similar developing countries around the world.

Keywords: Artificial Neural Networks, waste generation, environmental management.

ACKNOWLEDGEMENTS

Glory be to the almighty for His grace and sustenance. My sincere gratitude goes to my family for believing in me to make such investment into my education.

My profound gratitude goes to Prof. Roman Jašek for his contributions and advice to make this work easier. Furthermore, I am grateful to the lecturers at the faculty of Applied Informatics who in one way or the other imparted knowledge to me through teaching. Not forgetting my friends who helped me in diverse ways towards the success of this thesis. Finally, I dedicate this to my uncle Albert who passed away in May 2023, this one is for you.

God bless you all.

I hereby declare that the print version of my master's thesis and the electronic version of my thesis deposited in the IS/STAG system are identical.

TABLE OF CONTENTS

INTRODUCTION

The development of a sustainable environment is a matter of essence to the survival of humanity. Wendell Berry, an American environmental activist once said “the earth is what we all have in common” which is a fact. There are several methods for improving the condition of the environment that involve the implementation of integrated environmental management techniques. The main goal of environmental management is to implement solutions that monitor and examine the social, industrial, and economic effects on the environment in the long-term.

Environmental Management is a whole system which includes conservation, waste management, environmental law and policy, clean production, regulation, environmental knowledge, study of environmental impacts of human activities and environmental research.

The environment, which enables us to live by interacting with other things within it, is gradually deteriorating because of human activities, thereby contributing to environmental problems.

Environmental management problems consist of the three main types of pollutions. These are air pollution, land pollution and water pollution. Environmental pollution refers to the addition of substance of any form (solid, liquid, gas, heat, sound, radioactivity) to the environment at a rate faster than it can be decomposed, recycled, diluted, dispersed, or stored in a harmless form.

Environmental pollution primarily arises from the use of the environment by producers and consumers as a place to dispose of the waste products. The amount of waste and problems involved in disposing waste have increased as many countries have grown wealthier. Amid the many environmental management challenges faced today, one crucial area management of solid waste in municipal areas [1].

Municipal solid waste is understood as waste produced in municipalities. Most of these solid wastes are from households or from other sources, these solid wastes are often generated without any segregation, thus they may be either harmless or harmful according to their composition or properties. 2.01 billion tonnes of Municipal Solid Waste (MSW) are generated annually all over the world, with a projected increase of 3.40 billion tonnes by 2050 [2].

Most of the countries in the Middle East/North Africa are known for high MSW generation, with waste output exceeding 2 kg per capita per day on average. Solid waste is a global environmental management problem, especially in developing countries [2].

In Ghana, about 12,710 tons of solid waste is generated daily, with only 10% collected and disposed of at designated dumping sites. The lack of efficient solid waste management strategies is one reasons why environmental management issues abound in challenges faced by municipal authorities in Ghana [3].

Population growth, rise in community living standards and gradual economic development increases solid waste generation. Inappropriate disposal of solid waste can cause environmental contamination. Generated wastes which are not properly disposed often result in creating nuisance, aesthetic problems and pollution of land and water bodies of an area.

Proper planning of municipal solid waste management (MSWM) by way of forecast is necessary for building sustainable waste management system in the future. Planning and management of generated waste includes storage, collection, transfer, and disposal of waste. These are essential to any nation who considers environmental management to be a crucial factor for social wellbeing and healthy living.

An accurate prediction of the municipal solid waste generation rate is crucial for sustainable and efficient municipal solid waste management. Forecasting is a decision-making strategy that captures the trend of historical and current information to be utilized for future predictions. It has been used by many stakeholders such as academics, policymakers, government organizations, and municipalities to develop sustainable and effective municipal solid waste management systems. Therefore, creating a reliable municipal solid waste forecasting model for predicting generated waste quantity is very crucial, for a developing country such as Ghana.

One of such techniques is to create a model to forecast using Big data processing methods. Big data is a large volume of data which cannot be processed and stored using traditional computing approach within a specified time frame. Big data is available due to the abundance of information from numerous sources.

Machine learning, which is an example of big data processing methods is a multidisciplinary field that covers computer science, probability theory, statistics, approximate theory, complex algorithms. Machine Learning theories and methods have been widely used to solve complex problems in engineering applications. Machine learning also includes Support Vector Machine, Naïve Bayes, K-nearest neighbours, Artificial Neural Networks, etc.

This thesis focuses on the application of Artificial Neural networks which is a type of Big Data processing methods to predict the generation of solid waste in Ghana. Through this comprehensive review, the gaps, and emerging directions of Machine Learning application in Waste Management are thoroughly discussed, giving theoretical and practical guidance for follow-up affiliated research.

I. Problem Statement

It is estimated that 50 to 70% of the budget of municipal authorities is used to tackle management of waste. It has been reported that city authorities in Ghana spend about GHc 6.7 million (US$ 3.45 million) annually on the collection and transport of waste for disposal, and GHc 550,000.00 (US$ 0.28 million) per month to pay waste contractors and for landfill maintenance. Poor sanitation because of improper waste disposal alone is estimated to cost the country $290 million every year- an equivalent to 1.6% of the country's Gross Domestic Product [3]. Developing a waste generation prediction model with Machine learning algorithm namely Artificial Neural Network can help involved stakeholders develop proper implementation strategies that will save money, the environment and resources by proper planning and implementation.

II. Objectives

The aim of this study is to use machine learning algorithm namely Artificial Neural Network to create waste generation prediction model from statistical data obtained from agencies in Ghana. The results of this study are expected to represent a general outline for environmental management stakeholders in Ghana and these results can be extended to similar developing countries around the world.

III. Methodology

The prediction of MSW generation quantity in some selected cities Ghana was performed based on artificial neural network algorithm. MSW prediction using ANN is done using a series of processes. Data collected were fed into Neural Network library in MATLAB and Simulink software using a suitable architecture to get accuracy in the result. Neural network time series Toolbox (ntstool) was used for the study.

It is made up of three sequential levels including training, validation, and testing the result. The collected MSW data was used to train, validate, and test to forecast MSW generation in Ghana. The acceptable ratio of dataset was considered for training the model using the Levenberg–Marquardt algorithm. Statistical techniques for evaluating prediction result, i.e., mean square error (MSE) and regression coefficient (R), were measured for performance, final decision, and best model structure.

IV. Limitations

Limitations due to lack of inadequate data from related data collection agencies in Ghana.

I. THEORY

1 ENVIRONMENTAL MANAGEMENT

At the recently ended Conference of the Parties (COP26) UN Climate change conference in Glasgow, national leaders were compelled to take vital measures to lower emissions, protect and restore ecosystems and manage land sustainability [4]. The reason has been that an environment devoid of pollution provides an essential atmosphere that is crucial for human health.

The environment is a term used to describe nature or surroundings. Others see the environment as a natural landscape consisting of non- human features, processes, and characteristics. It is made up of the biotic and abiotic surrounding of an organism or population, it also consists of the factors that influence their survival, evolution, and development.

It can be a microscopic or global to an extent. Some examples of the environment include the atmosphere environment, terrestrial environment, and the marine environment. The Environment can be grouped into parts which include:

Atmosphere: which is the protective blanket of gases which surrounds the earth. It is very important for survival on this planet.

Hydrosphere: this is total amount of water on the planet. It consists of all kinds of water resources like the ocean, rivers, seas, lakes, streams, glaciers, ground water, etc.

Lithosphere: this is the solid outer part of the earth's surface. It consists of the upper section of the mantle and the crust of the earth surface.

Biosphere: this is defined as the realm where organisms live and interact with their surroundings. The biosphere is a very broad and complex term, and it is divided into smaller units known as ecosystems. Animals, plants, and microorganisms that reside in a particular zone with factors such as water, soil and air make an ecosystem.

Environmental management may be defined as the management of the interaction and impact of human activities on the natural environment. Environmental management focuses on marine, land, and atmospheric conditions, it also deals with issues associated with deforestation and global warming.

Humans are the primary agents that alter the conditions of the environment. This feat was achieved when humans discovered ways to tap into the energy supplies buried within the earth.

Energy obtained from burning fossil fuel provides power for our modern day businesses.

Environmental management is focused on understanding the functioning and structure of the earth and how best humans fit in. The ecological construct which creates a balance on our planet has been disrupted by industrialization, increase in population, transport, unplanned urbanization, technological revolution, and exploitation of natural resources.

The relationship can become better when individuals begin empowering exercises like preservation, recovery, and security of nature.

Environmental management deals with the prediction and maximization of human benefit and impact on the environments whiles reducing environmental degradation. Environmental management further seeks to ensure that the ecosystem is protected so that the future generation can use.

Environmental management involves:

- Identifying the environmental desired outcomes.
- Identifying the political, economic, cultural, social, and technological constraints on obtaining these outcomes.
- Considering the most feasible options for achieving the desired outcomes.
- Predicting, avoiding, and solving environmental and conservation issues.

With respect to its multidisciplinary nature, environmental management includes a diverse set of groups. For example, academics, policymakers, non-governmental organization workers, company employees, and civil servants.

The importance of environmental management is to reduce factors that contribute to climate change. Some of these factors include the following:

- Reducing carbon emissions.
- Processing all the waste in an effective and safe manner.
- The effective and wise use of energy and resources.
- Preventing pollution.

In 2015, ISO 14001, the world's leading environmental management system standard was substantially revised. The reason for the revised standard is to "Provide organizations with a framework to protect the environment and respond to changing environmental conditions in balance with socio-economic needs.

The revised standard makes clearer reference to the two-way relationship between organisations and the environment, which means that the organisation shall consider environmental conditions being affected by or capable of affecting the organisation (ISO 14001:2015 clause 4.1) [5].

It therefore discusses the requirement for organisations to address their impacts on the environment, including their contribution to climate change, but also now introduces the importance of resilience and adapting to our changing world.

1.1 Climate change and Solid waste management

Climate change remains one among the foremost important challenges facing humanity. It affects every country and disrupts national economies and affects lives, costing people, communities, and countries significantly today and within the future [6].

In addition, there's also a big inequity between countries' emissions and impacts, meaning that always those that contribute to global climate change the smallest amount, suffer from it the foremost.

People are experiencing the growing impacts of global climate change, which include changing weather patterns, rising water level, and more extreme weather events. Climate change impacts are one of several environmental impacts that derive from solid waste management options [6].

Other impacts include health effects attributable to air pollutants such as SO2, dioxins and fine particles, contamination of water bodies, emissions of ozone depleting substances, depletion of non-renewable resources, noise, accidents etc.

1.2 Waste

Wastes can be defined as substances or materials, other than radioactive materials covered by other international agreements, which: are disposed of or are being recovered, are intended to be disposed of or recovered, are required by the provisions of national law, to be disposed of or recovered [7].

Waste is sometimes by-product that is discarded as no longer useful after the completion of a process. It emanates from various sources such as agriculture, industry, households, and businesses. Waste can be in the form of liquid, solid, gaseous. It can also be hazardous or non- hazardous depending on the concentration and location.

1.2.1 Classification of waste

Waste generated by the society needs to be classified to apply the most efficient form of management. On a global scale a notable variety of definitions and classifications are used. Waste is classified based on substance (what it is made of), source (who or what generated the waste), hazard properties (how dangerous it can be), management (who manages it) or a mix of these concepts. The figure below shows the classification of waste by its origin, it further explains different activities that generate several types of waste.

Figure 1: Waste classification by origin [7]

Two major waste categories can be established based on the policy instruments and distinct legislation usually in place. These are non-hazardous or solid waste and hazardous waste. This classification is also used in the Basel convention. Usually, hazardous waste is managed at the national level whiles non- hazardous waste is managed at the regional or municipal level [7].

Non-hazardous/ Solid waste refers to waste that has not been classified as hazardous, for example paper, plastics, glass, beverage cans, metals, organic waste, etc. Solid waste can cause environmental and health impact if not managed properly.

Hazardous waste refers to waste that has been identified as potentially harmful to the environment and human health and therefore needs special treatment and handling. Chemical and physical characteristics determine the exact collection method and recycling process. Flammability, toxicity, and explosiveness, corrosiveness, are some of the characteristics of hazardous waste [7]. Incineration, Chemical treatment, safe storage, recovery, and recycling are feasible ways of treatment for hazardous waste.

Types of hazardous waste include.

- **Medical waste**: this originates from the human and animal healthcare systems. They usually comprise of medicines, pharmaceuticals, chemicals, used medical equipment, bodily fluids, and body parts. Medical waste can be toxic, infectious, radioactive or contain dangerous microorganisms.
- **E-waste:** This is waste from electronic and electric equipment such as computers, phones, and home appliances. E-waste is classified as hazardous because it usually contains toxic components.
- **Radioactive waste** is from radioactive materials. The management of radioactive waste differs significantly from that of other waste. Special treatment facilities are set up to handles this type of waste because of its toxicity.

Figure 2: Classification of waste [7]

1.2.2 Solid waste

Solid waste can be defined as a material that no longer has any use to the person who is responsible for it and is not meant to be disposed through a pipe. This type of waste does not usually include human excreta. It is usually generated by domestic, healthcare, industrial, agricultural, and mining activities which piles up in streets and public places. The words “trash” and ““garbage” is used to describe to some forms of solid waste.

Waste generated everyday consists of biodegradable organic matter such as, garden waste, kitchen waste and paper, which on average accounts for about 58% of the total weight of waste generated. In some bigger cities, the amount of organic waste sums up to about 70% of the total waste generated. Most of these rubbishes end up in landfills or in dumpsites [8].

When organic waste decomposes, methane gas and carbon dioxide are created. Methane is created when there is no air present while carbon dioxide is the natural product that is produced when anything rots in air. Carbon dioxide and methane are both greenhouse gases, which contribute to climate change and global warming.

All these factors need to be properly considered in the determination of a balanced policy for sustainable waste management, of which the climate change elements are but one aspect.

Improper management of waste represents a large amount of greenhouse gases, and some visible effects of climate change that can be seen include:

- Changing weather patterns.
- Increasing sea levels.
- Hotter temperatures.
- Increased drought

Waste management is more than collecting huge piles of trash and cleaning up the streets. Climate change has accelerated the need to discover measures to reduce and manage the waste we create. Management and reuse of waste can help reduce stress on the planet’s natural resources whereas possibly reducing emission of greenhouse gases made by mass production and burning of fossil fuel.

Even though governments must develop structures and policies to deal with increasing waste at the national level, communities and individuals can help to reduce the production of waste. It's a more difficult task and the implementation of big Data methods can help to understand the needs and adapt the practices most suited for environmental management.

2 BIG DATA

Data refers to a piece of information such as measurements or statistics. Big data is defined as complex and large data sets that are processed and thoroughly analysed to discover valuable information which can be useful to organisations and businesses.

In the year 2017 alone, we generated more data than in the previous 5,000 years. Big data is the technology that enables us to analyse this abundance in information and develop new advances and solutions. Application of big data in environmental management is also helping to optimise efficiency in the energy sector, to make businesses more sustainable and to create smart cities [9].

Data has rapidly ended up one of the foremost profitable assets for any organization. Modern tools and data-collection plans have made it conceivable to accumulate and analyse enormous sets of data dubbed "Big data."

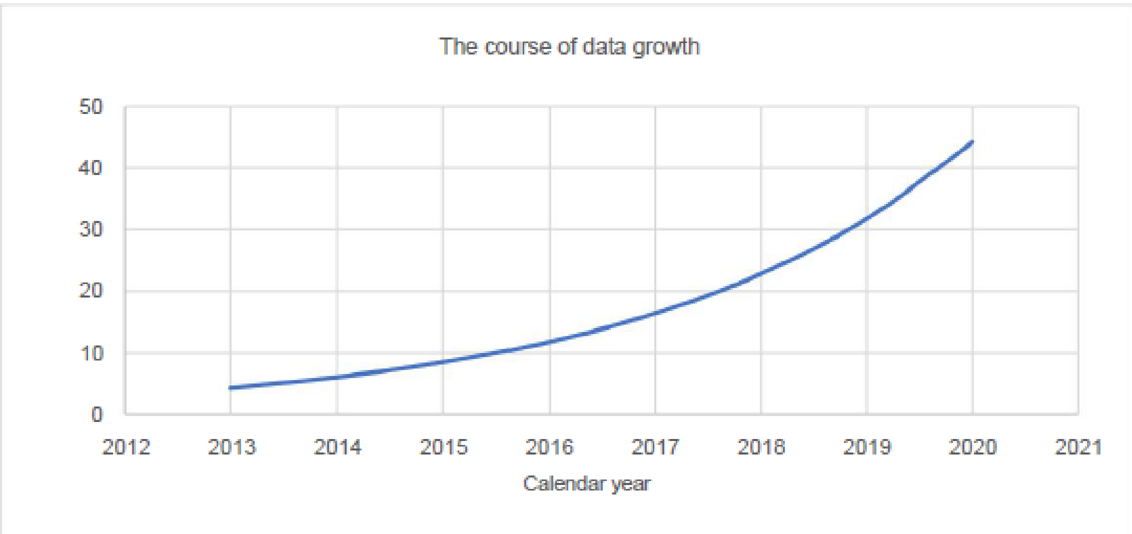

Figure 3: Predicted amount of data from 2013 to 2020 [9]

These data sets permit us to find designs and connections that we could not have found with conventional approaches. One sector being reshaped by enormous data is environmental protection and management. Governments, researchers, and scientists are being provided with new and excellent data conceivable, which allows them to make more accurate and data driven decisions.

The term Big Data has spread due to new technologies and innovations that have emerged over the past decade, the demand for the analysis of large amounts of data and rapidly generated data are on high levels and traditional computing approaches cannot solve them.

In the work of Dubey, he further stressed that Big data is also gaining ground in the field of sustainability, so it can be used to improve environmental and social sustainability in supply chains. It can also improve the allocation and utilization of natural resources [10].

Given the context of combating climate change, existing research has applied big data methods in fields such as of energy efficiency, smart urban planning, weather forecast, natural disaster management, intelligent agriculture, and waste management [11].

2.1 Big Data Characteristics

Big data is characterised by a set of parameters that explains different approach to big data analysis. The main characteristics of big data are classified by the 5V key characteristics, which are volume, velocity, variety, veracity, and value.

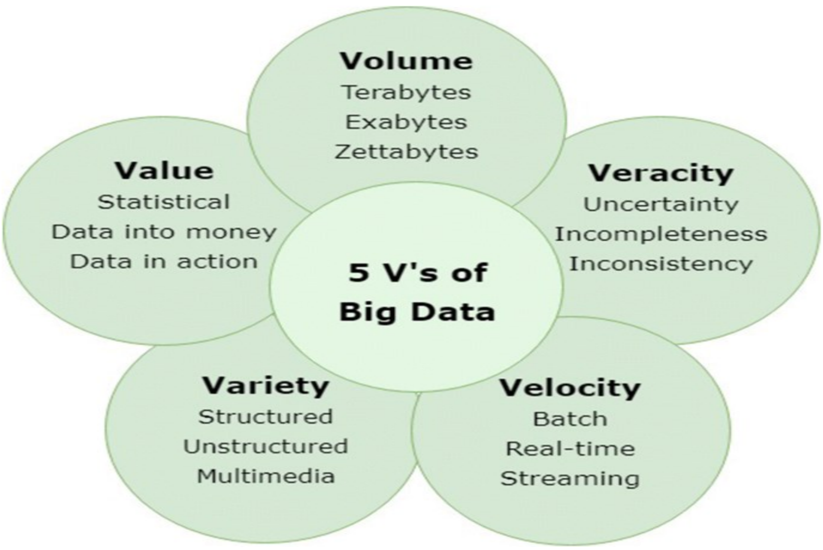

Figure 4: Characteristics of Big Data [12]

2.1.1 Volume

Volume refers to the huge amounts of information generated every second from social media, credit cards, images, cell phones, cars, video, etc. It is impractical to define a universal threshold for big data volume because the time and type of data can influence its description. Currently, datasets that reside in the exabyte or zettabyte ranges are generally considered as big data however challenges still exist for datasets in smaller size ranges [12].

2.1.2 Variety

This refers to the different forms of data in a dataset including structured data, semi-structured data, and unstructured data hat are being collected from different sources. Structured data is mostly well-organized and easily sorted, but unstructured data (e.g., text and multimedia content) is random and difficult to analyse. Semi-structured data (e.g., NoSQL databases) contains tags to separate data elements but enforcing this structure is left to the database user [12]. Big Data is basically categorised as below:

- **Structured Data**: It has a well-defined structure; it follows a consistent order, and it is designed in such a way that it can be easily used by a person or a computer. Structured data is usually stored in well-defined columns known as Databases. A typical example is the Database Management Systems.

- **Semi- Structured Data:** This can be considered as another form of Structured Data. It inherits a few properties of Structured Data, but a major part of this kind of data fails to have a definite structure and, it does not follow the formal structure of data models such as Relational Database Management System.

- **Unstructured Data:** This is a different type which neither has a structure nor follow the formal structural rules of data models. It does not have a coherent format and it varies all the time. But rarely it may have information that is related to data and time. An example is Audio Files, Images, etc.

- **Multimedia:** Multimedia is the presentation of text, pictures, audio, and video with links and tools that allow the user to navigate, engage, create, and communicate using a computer. Multimedia refers to the computer-supported integration of text, drawings, images, videos graphics, audio, animation, and any other media in which any type of information can be expressed, stored, and processed digitally.

2.1.3 Veracity

Veracity means how much the data is reliable. It has numerous ways to filter or translate the data. Veracity refers to handling and managing data efficiently. Big Data is also important in business development. For example, Facebook posts with hashtags.

2.1.4 Value

Value represents the context and usefulness of data for decision making, whereas the previous V's emphasis more on representing challenges in big data. For example, Google, Facebook, and Amazon have leveraged the value of big data via analytics in their respective products. Amazon analyses large datasets of users and their purchases to provide product recommendations, thereby increasing sales and user participation**.**

2.1.5 Velocity

Velocity comprises the speed (represented in terms of batch, real time, and streaming) of data processing, emphasizing that the speed with which the data is processed must meet the speed with which the data is produced. Velocity contains the linking of incoming data sets speeds, rate of change, and activity bursts. The main aspect of Big Data is to provide demanding data rapidly. Big data velocity involves speed at which data flows from sources like application logs, business processes, networks, and social media sites, sensors, mobile devices, etc.

2.2 Big Data Sources

Data sources refers to the origin of data. It may be the original location where the data is born or where the physical information is first digitized. Even the most refined data may serve as a source, if another process accesses and uses it. Basically, a data source may be a database, a file, live measurements from physical devices, web data, or any of the streaming data services available on the internet.

- **Sensors and activity records from electronic devices:**

This kind of information is produced on real-time, the number and periodicity of observations will be variable, occasionally it will depend on a lap of time, and on the occurrence of some event. In other instances, it will depend on manual manipulation. Example of these devices are medical devices, smart measuring devices, satellites, cameras etc.

- **Electronic Files:**

These refers to unstructured documents which are stored or published as electronic files, like Internet pages, videos, audios, PDF files, etc.

They can have contents of special interest but are difficult to extract, different techniques like pattern recognition, text mining, etc could be used.

- **Social Networks:** Twitter, Facebook, Instagram, YouTube, etc .
- **Data Storage:** NoSQL, file systems, SQL etc.
- **Media:** Images, audio, podcast, video, etc.

2.3 Applications of Big Data

The concept of big data approach is currently enormous. Big Data is considered a powerful asset that can run big IT industries of the 21st Century. Big Data finds application in all areas of human activities.

It is one of the most common technologies that is being used in almost every business sector. A few of these sectors are mentioned below:

2.3.1 Healthcare

Healthcare is another sector that generates large amount of data. Listed below are some of the ways in which big data has contributed to healthcare:

- Big data reduces the costs for treatment since there are lower chances of having to perform unnecessary diagnoses.
- It helps in predicting outbreaks of epidemics and in deciding what preventive measures could be taken to minimize the effects of the same.
- It helps avoid diseases by detecting them in the early stages. It prevents them from getting any worse thus making their treatment efficient.
- Patients can be given evidence-based medicine from researching past medical results.

2.3.2 Transport Industry

Big data has been used in various ways to make transportation more effective and easier. Listed below are some of the areas where big data contributes to transportation.

- **Route planning**: Big data can be used to understand and predict users' needs on different routes and then utilize route planning to reduce their waiting time.

- **Congestion management and traffic control**: The use of big data has now made real-time estimation of congestion and traffic patterns possible. For example, the use of Google Maps to locate the least traffic-prone routes.
- **The level of traffic**: The use of real-time processing of big data and predictive analysis to identify accident-prone areas can help reduce accidents and increase the safety level of traffic.

2.3.3 Marketing

Traditional marketing approaches were based on the survey and interactions with the customers. Companies would run advertisements on TV channels, radio stations, newspapers, and mount banners on billboards by the roads. These activities went a long way to impact customers on their buying preferences.

With the evolution of the internet and technologies like big data, the field of marketing also went digital, which is now known as Digital Marketing . Today, with big data large amount of data can be collected to know the preferences of millions of customers in a few seconds. Business Analysts analyse the data to help marketers improve the product, run campaigns put relevant advertisements.

2.3.4 Environmental Management

Environmental agencies have employed these Big data to improve their practices and refine their methods. Weather sensors and satellites are stationed all around the world. A large amount of data is collected from them, this data collected is then used to monitor the environmental conditions and weather. An example of such technologies is the Aqueduct's mapping tool which allows users to monitor and calculate water risk anywhere in the world, all from the comfort and convenience of their laptop.

Data collected from these satellites and sensors contribute to big data and can be used in several ways such as:

- weather forecasting.
- studying global warming.
- understanding the patterns of natural disasters.
- making necessary preparations in the case of crises.

- predict the availability of usable water around the world.

Countries around the world can keep track of their emissions, reaching renewable energy goals as they raise standards of sustainability in all sectors, such as agriculture.

2.3.5 Government Sector

Governments, regardless of the country, process huge amount of data on daily basis. The reason for this is that they must track and monitor various records and databases regarding their citizens, their growth, energy resources, geographical surveys, etc. All this data contributes to big data. The proper study and analysis of this data helps governments in numerous ways. A few of them are listed below:

- Making faster and informed decisions concerning various political programs.
- To overcome national challenges such as, terrorism, energy resources exploration, unemployment, etc.
- Big Data is used for fraud recognition in the domain of cyber security.
- Big data is also used in catching tax evaders.
- Cyber security engineers protect networks and data from unauthorized access.

2.3.6 Media and entertainment

The media and entertainment industries are creating, advertising, and distributing their content using new business models. This is as the result of customer requirements to view digital content from any location around the world and at any time. The introduction of online TV shows, HBO, Netflix channels etc. is asserting the fact that new customers are not only interested in watching TV but are also interested in accessing data from any location.

The media industries are targeting audiences by predicting what they would like to see, content monetization, how to target the ads, etc. Big data systems are thus increasing the revenues of such media houses by analysing viewer patterns.

3 BIG DATA PROCESSING METHODS

Big Data methods affect the method of examining data sets (within the shape of text, audio, and video), and drawing conclusions about the knowledge they contain, more commonly through specific systems, software, and methods. Big Data methods are technologies that are used on an industrial scale, across commercial business industries, as they allow organizations to form calculated, informed business decisions.

Big data has evolved as a product of our increasing expansion and connection, and with it, new sorts of extracting, or rather "mining" data. As the generation of knowledge increases, so will the varied techniques that manage it.

McKinsey's big data report identifies a variety of massive data techniques and technologies, that draw from various fields like statistics, computing, applied math, and economics [13]. Types of Big Data processing methods are explained below:

3.1 Data fusion

Data fusion involves integration multiple sources of data to create more accurate, consistent, and useful information. Data fusion involves sensors that collect sets of information. These sensors can be any type of sensors that can collect thermal, visual, audible, and other types of input. Once the data has been collected, data fusion algorithms integrate the data into a single data set which is more understandable [14].

Data fusion are usually grouped as low, intermediate, or high depending on the stage at which the fusion occurs. In low level fusion several sources of raw data are combined to produce new data. The idea is that the fused data is more informative that the original inputs. By combining a group of techniques that analyse and integrate data from multiple sources and solutions, the insights are more efficient and potentially more accurate than if developed through one source of knowledge.

3.2 Data processing

Data processing is converting data into a usable and desired form. The processing it usually done by computers automatically. The output of the processed data can be available in various forms such as audio, charts, table, images, etc. In data processing datasets are fed into a special software to filter out useful information from it [15].

Data processing is a common tool used within big data analytics, data processing extracts patterns from large data sets by combining methods from statistics and machine learning, within management.

The six stages of data processing include:

- **Data collection**: in this stage data is collected from available data sources. It is important that the data collected are credible in order get quality information after processing.
- **Data preparation:** this stage comes rightly after the data is collected. In data preparation the raw data is organised and check for error. This is done to remove bad or incorrect data to create high quality results.
- **Data input:** in this stage data is fed into a special software to be translated to a language that the software can work on.
- **Processing:** in this stage data is processed by special computer software and machine learning algorithms. Data processing varies depending on the type of data collected.
- **Data output:** in this stage the processed data becomes available and usable to users. It is usually displayed in the form of images, graphs, charts, videos, plain text, etc so that users can understand.
- **Data Storage:** this is the final stage of data processing. after the data has been processed it is passed on to a storage for future use and reference.

3.3 Statistics

Statistics refers to collecting, analysing, and understanding data and accounting for relevant uncertainties. Statistics if essential to extracting meaningful and useful information from large data sets. Statistical methods measure uncertainty, design studies, create sampling strategies, access quality of data. Statistical methods are sometimes used to create models for analysing complex data structures. This technique works to gather, organise, and interpret data, within surveys and experiments. Other statistical techniques include spatial analysis, predictive modelling, association rule learning, network analysis, etc. The technologies that process, manage, and analyse this data are of a completely different and expansive field, that similarly evolves and develops over time.

3.4 Machine learning

Machine Learning is a field that covers statistics, computer science, probability theory and complex algorithms. Its theories and methods are used to solve complex problems in the field of engineering. These algorithms find hidden information from a large pool of historical data and uses them for either classification or regression. Machine learning also includes Support Vector Machine, Naïve Bayes, K-nearest neighbours, Decision Trees, Artificial Neural Networks, Random Forests and Fuzzy inference systems. Machine learning algorithms have three main methods of learning namely supervised, unsupervised and reinforcement learning. These methods of learning are explained below:

- **Supervised learning**: this is one of the simplest types of machine learning. In this type of learning input and output data are required. The machine learning algorithm is trained on actual labelled data. Applications of supervised learning includes image classification, time series prediction and categorization. Examples of supervised machine learning algorithms are Support Vector Mechanics, K Nearest Neighbour, Naïve Bayes, Decision tree.
- **Unsupervised learning:** this type of machine learning only requires input data. One major advantage us that it can work with unlabelled data. Meaning that it does not entirely require human efforts to make the data set machine readable therefore allowing much larger datasets to be worked on by the algorithm. Unsupervised learning creates hidden structures because they have no labels to work with. The creation of these hidden structures makes unsupervised learning algorithms versatile. This is since relationships between data points are perceived by the algorithm in an abstract manner, with no input needed from human beings. Examples of unsupervised learning are K means Clustering, DBSCAN, Hierarchical Clustering.
- **Reinforcement learning:** this type of leaning takes inspiration from how human beings can learn from data in their lives. It does not require input and output labels Reinforcement algorithms try to improve themselves and learns from new situations using a trial and error method. Favourable outputs are reinforced, and unfavourable outputs are ignored. Examples of reinforcement leaning includes Monte Carlo, Deep Q Network, Sarsa, Q- Learning.

3.4.1 Linear regression (LR) Machine Learning approach

Linear Regression is described as one of the most recognized and well-understood algorithms in machine learning. Predictive modelling with Linear Regression mostly focusses on minimalizing the error of a model and making the most precise forecasts possible [16].

Once a statistical relationship is established between the input and output variables in solid waste management, Linear Regression approach becomes a most frequently used model. Multiple Linear Regression (MLR) and Single Linear Regression (SLR) have been used in MSW prediction process. Level of education, urban morphology, tourism activity, and income as the influencing factors that have mostly led to MSW generation [17].

Single Linear Regression has been used to link gross domestic product (GDP) and related total consumer expenditure as very strong correlating factors in MSW generation at the country level. However, the limitation of the LR models is generally associated with low coefficient of correlation value [18].

3.4.2 Artificial neural network (ANN) ML approach

ANNs are named after neurons in the human brain. They are made up of algorithms modelled like the human nervous system and are designed to recognize patterns and relationships in data. They function like neurons in the human body, they receive stimuli, work on them, and transmit them to other processing units. Artificial neural networks are complex data handling systems designed and built on the concept of neural system in human [16].

One key feature of ANN in predicting MSW generation is in its ability to learn. ANN can construct a non-linear structure through a set of input/output data. Abbasi in his study discussed that the non-linear nature of MSW generated makes ANN one of the ideal Machine Learning tools in forecasting MSW production data [19].

A study by Ordonez-Ponce also utilized key factors such as socio-demographic, economic, geographic, and waste-related factors, a coefficient of determination (R^2) value of 0.819 were obtained from the results. They concluded that populace, section of city residents, level of education etc. were the significant variables contributing to MSW proliferation in Chile [20].

Also, Jalali in his study used ANN to predict MSW production in Mashhad, Iran. The result showed a very close correlation between the empirical data and prediction data. The ability of ANN in the short-term prediction of MSW generation was also observed in other Nouri studies conducted by them [21].

The authors concentrated on MSW generation forecast by studying the time series of waste generation instead of analysing the active factors in MSW generation. Their results showed that feed-forward ANN technique with one hidden layer and 16 neurons was the model that gave the best result to estimate short-term waste generation rate [16].

ANN may find it difficult to model non-stationary data precisely if pre-processing of the input variables and output variables is not carried out [19].

3.4.3 Decision tree ML approach

Decision Trees are a non-parametric supervised learning method used for classification and regression. The aim is to make a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. Decision Trees are non-parametric models which allow them to be used with different data types. It also requires minimum data transformations.

DT models may build complex trees with large dataset which can cause overfitting when training data with noise is supplied. Decision Trees have restricted ability in handling non-linear dataset when compared with other algorithms such as ANN [22].

Kannangara et al used classification and regression tree algorithm to create a DT model for MSW generation forecasting. From their results, DT models did not perform well in describing the variation in their dataset [16].

3.4.4 Supported vector machine (SVM) ML approach

SVM is another machine learning technique which can deliver an effective method that can simultaneously improve generalization performance of ANN model and achieve global solutions. Support vectors are a small subset of training observations which are used as support for the optimum location of decision surfaces.

SVM is a form of maximum-margin classifier that try to find a maximum margin hyperplane to the best fit across data. The hyperplane obtained in the process is referred to as the optimal separating hyperplane and the training examples that are closest to the maximum margin hyperplane are called support vectors [23].

Abbasi in his work probed further to model the uncertainty of the results obtained by using SVM and the hybrid of wavelet transforms in MSW generation forecast using the Monte

Carlo method. The results showed that both ML models had a precise prediction of MSW generation in both Tehran and Mashhab cities of Iran. They also realised that the use of wavelet transforms to pre-process the input variables led to the development of a more accurate SVM model in the weekly forecast of MSW generation in both cities [19].

2.4.5 Gradient Boosted Regression Tree (GBRT) ML approach

Gradient Boosted Regression Tree is an ensemble ML method which consists of multiple weaker models that are trained separately. Ensemble models create predictions which are further combined in different ways to make overall prediction.

It can deal with non-linear dataset that has complex relationship and as well as ease in interpretation, GBRT is a decision tree-based model that is often employed in ML. They have demonstrated a higher prediction accuracy than the regular DT models.

GBRT has become popular in machine learning due to its ability to handle complex non-linear relationships and its robustness to outliers. It has been successfully applied in various domains, including regression problems with continuous target variables.

Note that GBRT can also be adapted for classification tasks, known as Gradient Boosted Trees or Gradient Boosting Machines (GBM), where decision trees are used to predict class probabilities or labels instead of continuous values.

II. ANALYSIS

4 MATERIALS AND METHODS

4.1 The Case study area

Ghana is a country of West Africa with a population of 31 million and has a total area of 238,533sq. km with a coastal line of 550 km. It has a tropical climate with two major seasons: rainy season (May–October) and dry season (November–April). Ghana has three major geographic regions known as northern savanna, coastal and forest with no distinctly specified boundaries [24].

As of 2020, there are currently sixteen regions, which are further divided for administrative purposes into metropolitan, municipal and district assemblies (MMDA).

A major challenge in the management of solid waste in Ghana is the collection and disposal process, which are labour-intensive and often not effective. Issues relating to proper solid waste disposal in urban cities across Ghana is a major challenge for the local government authorities. City authorities and waste management companies are often overwhelmed by the quantity of waste generated daily [24].

Key factors impeding proper management of solid waste in Ghana are rapid population inadequate supply of waste bins, growth and urbanization, lack of waste transportation systems, low public awareness on the health consequences of poor waste management.

Also, weak enforcement of environmental regulations which are further divided into metropolitan, municipal and district assemblies, all of them having governing authorities [24].

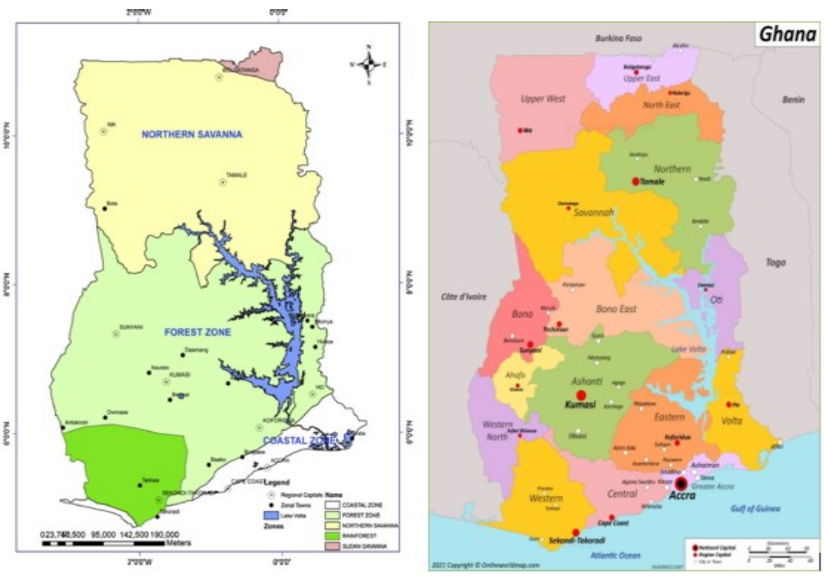

Figure 5. Map of Ghana [24]

4.2 Data Sets

The accurate measurement of solid waste is a demanding task for any government and authority. Therefore, the actual data availability on solid waste generation and management far behind, especially in a developing country such as Ghana. The dataset (Table 1) used in this project is acquired from previous research in Ghana by Kodwo Miezah [24].

In Table 1 the writer focused on generating a comprehensive data on the production of household's waste at the regional and national level in Ghana for planning and waste management activities [24]. Variables such as population, number of employed, number of unemployed, were based on the estimation done by the writers in relation to the statistics available by Ghana Statistics Service on 2010 Population and housing census. Percentage increase was calculated on the population used for the study and then it was applied on the other input variables.

Table 2 contains data about total waste generation by some sectors in Ghana from 1991-2015. The data was also obtained from a national population census conducted by Ghana statistical service [25].

Table 1. Analysed socio-economic factors and waste generation of households in Ghana.

Cities	2014 Population estimate based on 2010 census growth rate	Number of employed	Number of unemployed	Number of households	Total waste generation based on Population/tonnes
Accra	2,088,723	969,497	75,442	187,770	1552
Kumasi	2,263,914	976,424	91,784	611,209	1689
Takoradi	605,673	223,011	27,554	154,307	424
Cape Coast	191,961	75,135	6,968	45,628	128
Koforidua	199,653	85,109	6,704	53,763	122
Sunyani	134,958	53,169	3,906	26,342	66
Tamale	416,338	155,205	12,475	36,059	137

Wa	115,627	37,726	3,510	6,248	29
Bolgatanga	137,979	33,213	805	28,009	29
Ho	300,103	124,945	8,567	84,345	94

Table 2. Total Generated waste by sectors in Ghana from 1999-2015 – historical data

Year	Population estimate	Agriculture, fishing, and forestry [t]	Other Economic Activities [t]	Households [t]	Total Waste Generation [t]
1999	18,812,359	6,772.57	300.77	1203.08	8,276.42
2000	19,278,850	6,784.24	331.34	1325.36	8,440.94
2001	19,756,929	7,032.57	345.96	1383.83	8,762.36
2002	20,246,376	8,249.63	361.12	1444.49	10,055.24
2003	20,750,308	8,244.81	370.87	1483.49	10,099.18
2004	21,272,328	6,772.57	387.03	1548.12	8,707.72
2005	21,814,648	8,060.92	454.26	1817.05	10,332.23
2006	22,379,057	8,224.60	473.82	1895.27	10,593.68
2007	22,963,946	10,833.10	516.56	2066.22	13,415.88
2008	23,563,832	10,954.26	561.26	2245.02	13,760.54
2009	24,170,943	13,625.67	600.1	2400.39	16,626.16
2010	24,779,614	12,422.40	662.61	2650.46	15,735.48
2011	25,387,713	14,157.82	679.18	2716.72	17,553.72
2012	25,996,454	14,953.13	696.16	2784.64	18,433.93
2013	26,607,641	15,107.61	713.56	2854.25	18,675.43
2014	27,224,480	15,516.54	731.4	2925.61	19,173.56
2015	27,849,203	15,834.55	783.01	3132.03	19,749.59

4.2.1 Statistical Analysis of Data Sets.

The explanatory variables that contribute to the quantity and type of waste generation were carefully chosen based on correlation test and literature reviewed. The datasets were analysed using Pearson's correlation analysis to ascertain the relation between the input data (socio-economic factors for the chosen cities) and the output data. Pearson's correlation is calculated using this formula:

$$r = \frac{\sum(x_i - \overline{x}) - (y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \ \sum(y_i - \overline{y})^2}}$$

Where x_i are values of the x- variable in a sample, $\overline{x}$ is the mean of the values of the x-variable, $\overline{y}$ is the mean of the variable and y_i are values of the y- variable in a sample. It is assumed that correlation coefficient r > 0.7 shows that there is a strong relationship between the two parameters, values that range between 0.5-0.7 shows a moderate relationship between them and 0 means there is no correlation at all.

Table 3. Correlation analysis between input variables and output variables

Input Variables	Total waste generation based on population
Population	0.9967
Number of employed	0.9957
Number of unemployed	0.9952
No of households	0.8527

From table 3 it is observed that the correlation coefficients between the input variables (population, number of employed, number of unemployed and number of households) and the output (total waste generation) are positive. Meaning that if the various variables of the input increases, then the amount of waste with also increase. To interpret the results from Table 3. it is observed that the highest value for the correlation coefficient was 0.9967 and the lowest was 0.8527 which however indicates a moderate relationship between the two parameters.

Table 4. Statistical parameters of input and output variables for MSW prediction

Statistical Parameters	Population	Number of employed	Number of unemployed	Number of Household	Total MSW [t]
Mean	645,493	273,343	23,772	123,368	427
Median	249,878	105,027	7,768	49,696	125
Standard Deviation	821,856.3	373,249.68	32,617.2	181,154.4	639.7
Min	115,627	33,213	805	6,248	29
Max	2,263,915	976,424	91,784	611,209	1,689

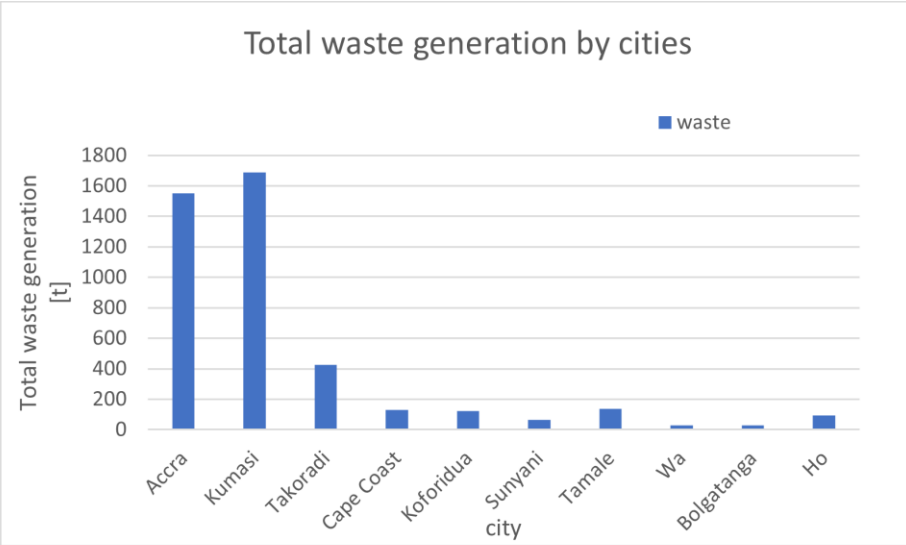

Figure 6. Distribution of municipal waste in Ghanaian cities.

4.3 Decision of input parameters

There are several socioeconomic factors that affect MSW generation. After a comprehensive review of various literatures certain features emerged as standout factors that directly influence waste generation in urban or municipal areas. The quantity of MSW depends on the general socioeconomic structures of countries.

Even in the same country, the waste generation and characteristics vary according to the population, lifestyles, socioeconomic status, consumption habits and traditions of the people living in different cities. These parameters were chosen as feature vectors since they are generic indicators of sustainability for countries with different level of economic development, productivity, industrial structure, and output.

Population: Population has been established as a key factor that contributes to acute MSW generation [11]. The volume of MSW is found to be strongly correlated with population in related studies. This is because rapid population growth increases the amount of waste generated.

People employed: The number of people employed in an area influences the generation of waste. When the income level of the people is high more solid waste is produced.

People unemployed: The number of unemployed in an area also influences the generation of waste since they have lower purchasing power.

Number of households: Household wastes are mostly generated from places where there is human interaction or activities. Several studies indicate that most of the municipal solid waste from developing countries are generated from households (55-80%) [24].

4.4 Machine Learning Method applied

Artificial Neural Networks, with its inspiration from neurons in the human brain possess a powerful pattern recognition and classification abilities. ANN can learn from experience, also they are able to capture relationships between input and output variables using cross-sectional data. ANN models make predictions of the future based on previous data. ANN's are generally made up of three components known as input layer, hidden layer, and output layer.

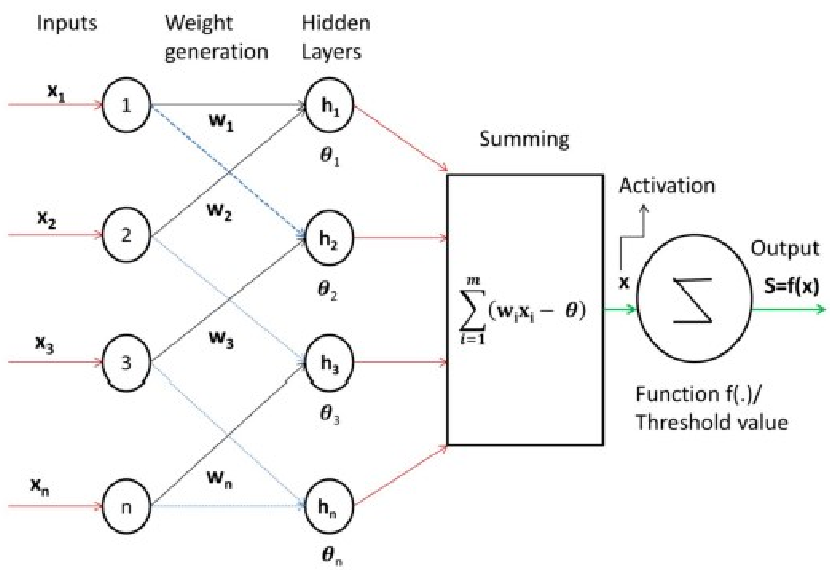

Figure 7. A diagrammatic representation of an artificial neural network [26]

Figure 7 shows a simple representation of an ANN. The input layer (input node) is the input or information that is fed to the model to learn. The number of neurons in the input layer are created based on the number of input variables while the number of different outcomes determines the number of neurons in the output layers. The hidden layer is made up of a set of neurons which performs computations on the input data.

The output layer is the final model obtained from the computations performed. $x_{1,}$ $x_{n,}$ are inputs whereas $w_{1,}w_{n,}$ are weights. The weights are the gradient of each variable. It signifies the strength of a particular input. The weights are then multiplied to their respective inputs and a bias is variable added. It is calculated using this formula:

S =$w_{1,}\ x_{1,} + w_{2,}\ x_{2,} + \cdots .,= \ \sum w_{i,}\ x_{n,}$

S = $\sum_{i-1}^{n}(w_{i,}\ x_{i,-\theta})$

Where $x_{1,} \ldots x_{n}$are inputs neurons, $w_{1} \ldots . w_{n,}$ are the weights, θ is the bias term, S is the Summing function and x is the activation function. The weights and biases are assigned to channels linking neurons in following layers. The summing function measure the threshold value. The threshold value is used to compare the difference between the input and output. If the weight of output sum S ($f(x)$) is greater than this threshold value, then the output will be 1 otherwise 0.

The activation state of the neuron is then determined by passing this value through an activation function. The data are propagated forward through activated neurons in the different layers until the result is delivered in the last layer [27].

There are three threshold functions that are generally used i.e., sigmoid function, piecewise linear function, and the hyperbolic tangent function. The sigmoid function was employed in this study because besides it being differentiable it can compress values within the range of 0 and 1 therefore making it easy to work with.

The neural network is trained by comparing the expected output to the real result. The weights are then adjusted based on the computed prediction error in the back-propagation process. The forward and backward propagation procedures are repeated until the network can accurately predict the output.

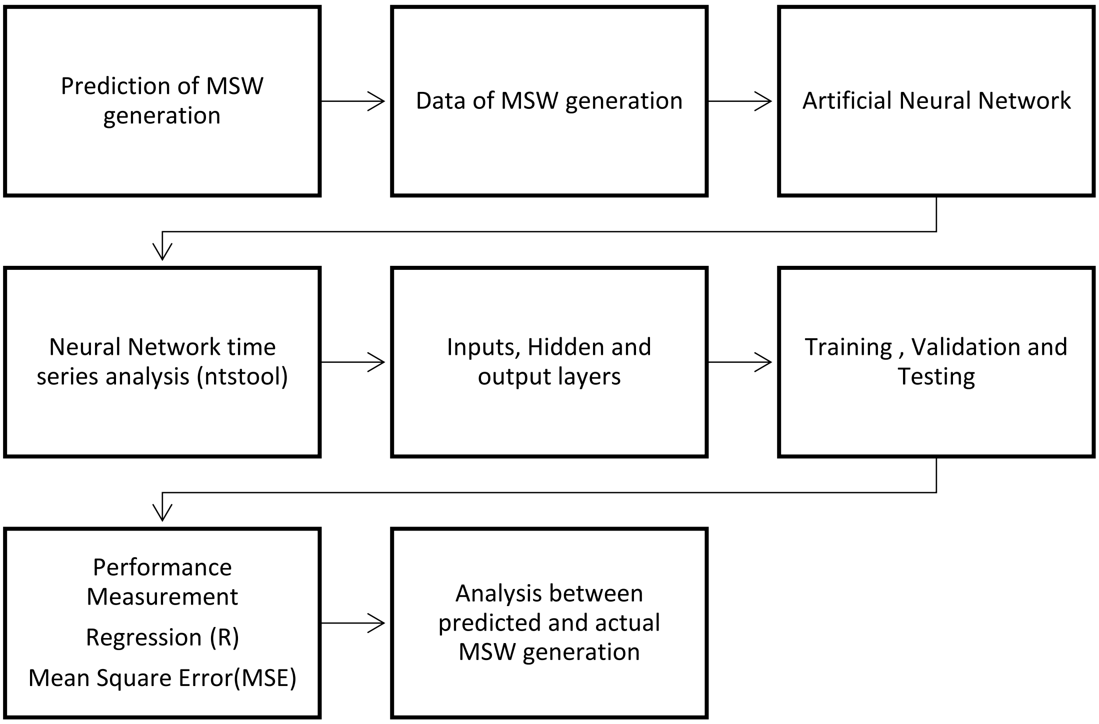

Figure 8: A chart showing the procedure used for the study.

4.5 Evaluation metrics

There are numerous standard techniques that can be used to measure the performance of forecasting models in a numerical study To evaluate the prediction performance of the ANN model, two statistical metrics were used: Mean Square Error (MSE) and regression value (R). MSE is calculated with this formula:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2$$

Where n is the number of data points in a dataset, Y_i is the observed values, $\widehat{Y_i}$ is the predicted values. The regression value R is used to measure the correlation between the input and output variables. It shows how the well the input and output match. The trained network is considered good is the R value is close to 1. Regression value (R) is calculated with this formula:

$$\text{R}\,(y', y^*) = \frac{cov((y', y^*)}{dy'\,dy^*}\ \text{R}\in \,< 0{,}1 >$$

Where dy' is the standard deviation of the reference values, dy^* is the standard deviation of the predicted values. For generated result to be quality the regression (R) value must be higher and the MSE must be a lower.

4.6 MATLAB software

MATLAB is an abbreviation for Matrix Laboratory. It is a programming language which was created by MathWorks in Massachusetts, USA. It is a high performance language used for technical computing. It integrates computation, visualisation, and programming in an easy environment. The MATLAB environment makes it possible to perform technical calculations, algorithm design, simulation, data presentation measurements, signal processing, etc. It is an interactive system which allows you to solve many technical computing problems.

MATLAB is a tool that has an interactive interface, and it can be used for the development of applications. It provides uses with a programming language based on technical and mathematical computations. MATLAB is especially used in engineering field.

Some examples of the usage include:

- Machine learning
- Data Science
- Simulation
- Creating charts for Big Data
- Data Analysis and visualisation
- Algorithm Development

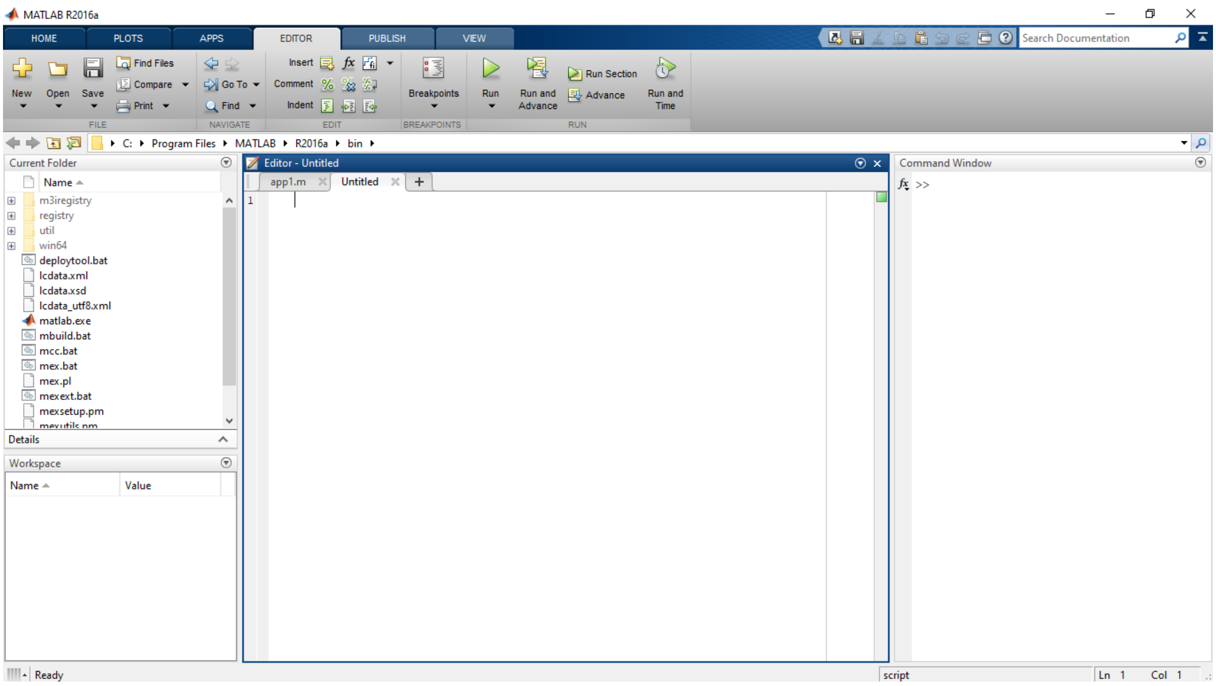

Figure 9. A simple MATLAB software Interface

4.6.1 Modelling of ANN in MATLAB Toolbox

The Neural Network Time Series Toolbox (ntstool) in MATLAB R2016 version was used for the ANN modelling. In this toolbox prediction is a dynamic filtering where past values of one or more time series are used in the prediction of future values. The neural network toolbox (ntstool) makes it possible to make predictive models which are used for system identification.

In this study a multilayer feedforward neural network was used with a nonlinear auto regressive technique (NARX) which is available in MATLAB software environment. There are two methods of forecasting in NARX neural network namely Open loop and closed loop forecasting.

Open loop forecasting makes prediction of the next time step in sequence with the use of the input data only. To make predictions for time steps, true values from a dataset are collected then used as the input. Open loop forecasting is used to make prediction when true values are available for the network.

Closed loop forecasting makes prediction of subsequent time steps in a sequence by using past predictions as the input. The model does not need true values to make prediction. Close loop forecasting is used to forecast multiple time steps or when true values are not available to the network to make the next prediction.

The MATLAB version 2016 offers three types of tools for solving time series problems in ntstool which are Nonlinear Autoregressive with External Input (NARX), Nonlinear autoregressive (NAR) and Nonlinear Input Output. NARX tool was used because it gives better accuracy as compared to the other tools. NARX tool predicts series **y(t)** given **d** past values of **y(t)** and another series **x(t)**. It is expressed by this formula.

$y(t) = f(x(t-1), \dots x(t-d), y(t-1) \dots y(t-d))$

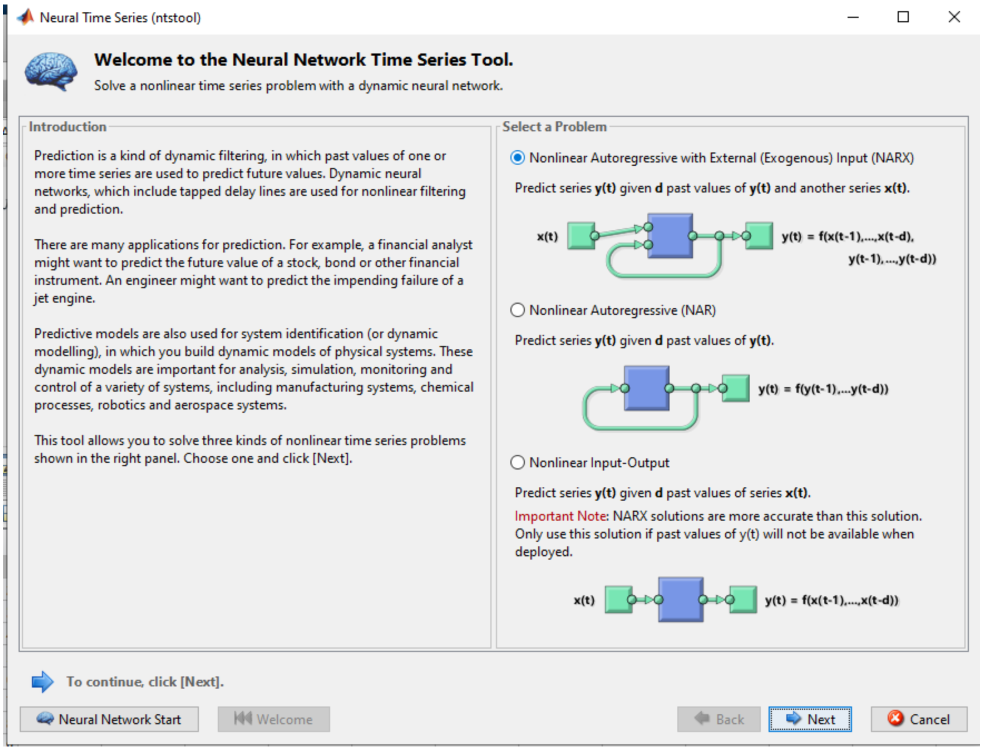

Figure 10. Diagram of a neural network Time series tool interface

4.6.2 Building the Network

In building the ANN the number of hidden layers, neurons in each layer, transfer function, training algorithm, and other parameters associated with setting the network were chosen. Two different datasets were used, and their results were analysed respectively. Dataset one (Table 1.) consists of socio-economic factors that contributes to waste generation in some selected cities in Ghana and the total waste the cities produced based on these factors.

The variables involved in waste generation modelling for the selected cities in Ghana were grouped into four parts (population, number of employed, number of unemployed and household) which were fed into the neural network as the input neurons x(t). The remaining variable (total waste generation) was then labelled as the output neuron y(t).

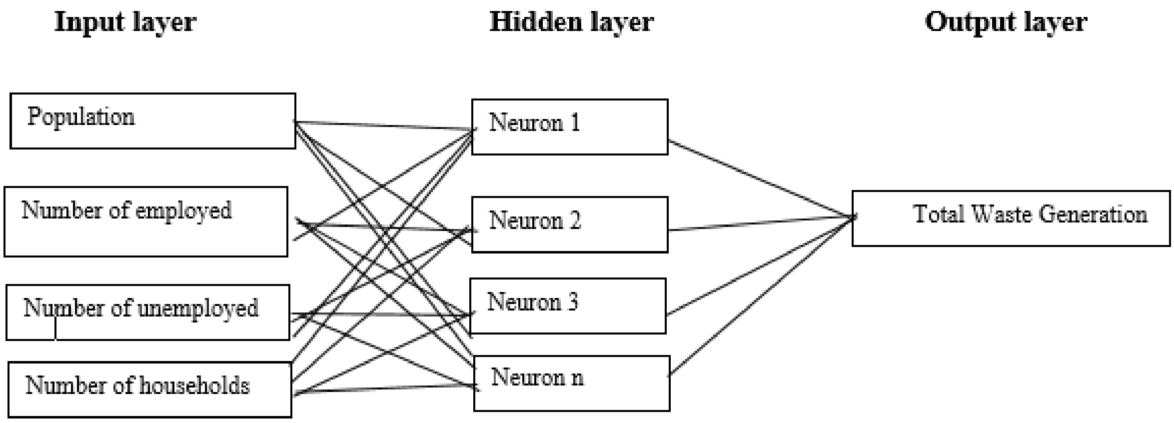

Figure 11. Neural Network structure based on Table 1

The dataset defined in Table 2 consists of the historical data of waste generated by some sectors in Ghana (Agriculture, other activities, Households) and their corresponding total waste generated from 1999 to 2015. The variables involved in waste generation modelling for the total waste generated on historic data in Ghana were grouped into two parts (year and population) which were fed into the neural network as the input neurons x(t). The remaining variable (Agricultural, forestry and fishing, other economic activities, households, total waste generation) was then labelled as the output neuron y(t).

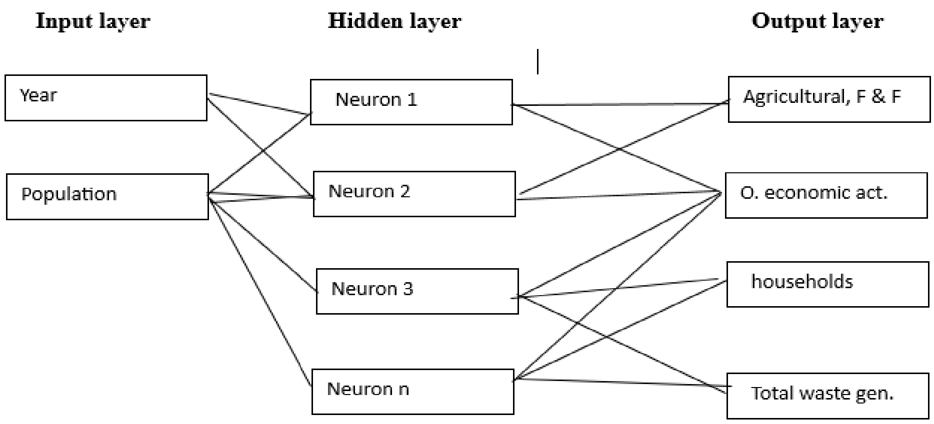

Figure 12. Neural Network structure based on Table 2

Levenberg- Marquardt algorithm was used for the training. The number of hidden layer neurons (2-10) were selected based on the experiment. Figure 10. below shows the diagrammatic representation of the selection and loading on data set in the MATLAB ANN toolbox environment.

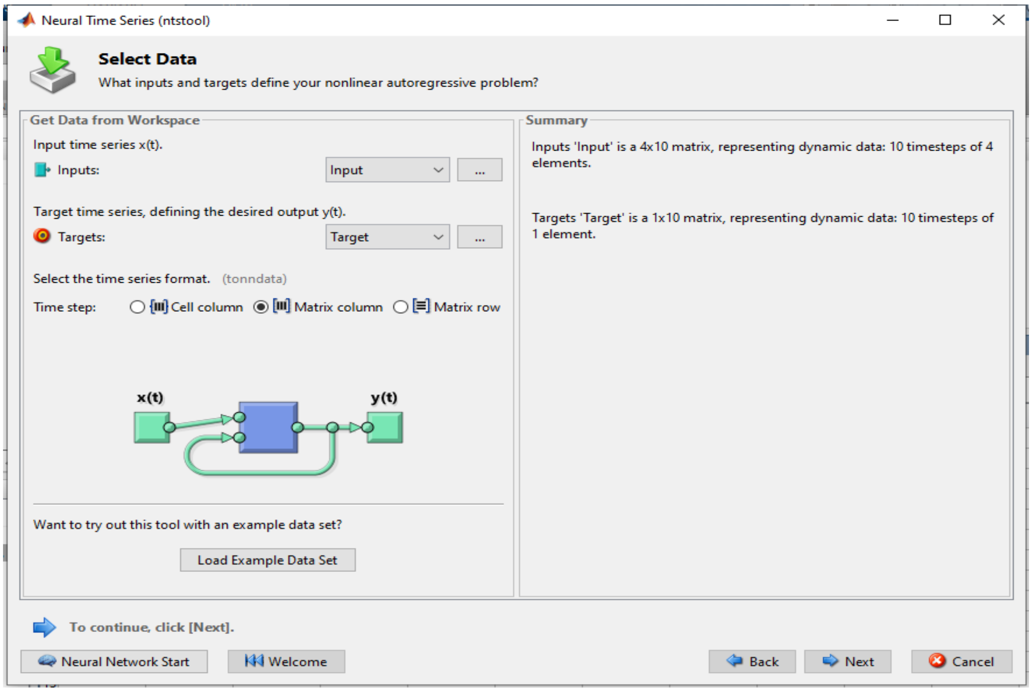

Figure 13. Diagram of data selection phase in MATLAB ANN toolbox

4.6.3 Training of the ANN Network

To train the data, MSW generation for the selected cities in Accra were fed into the MATLAB ANN toolbox and the data were randomly divided into three parts which are training (70%), validating (15%) and testing (15%) to avoid overfitting. The Levenberg Marquardt algorithm was used to train the dataset because it is a simple and robust algorithm which takes less time to train. The target time series dataset was selected as the desired output y(t). During the training, the weights were adjusted to make the predicted output close to the target output of the neural network. In training the ANN model learns from the previous data then uses the knowledge gained to make the future prediction. To obtain the model The neurons of the hidden layers were changed accordingly during the experiment and the best ANN structure is the one with the lowest value for the Mean Square Error (MSE) and the highest value of the regression (R). Figure 13. below shows the training phase of the network.

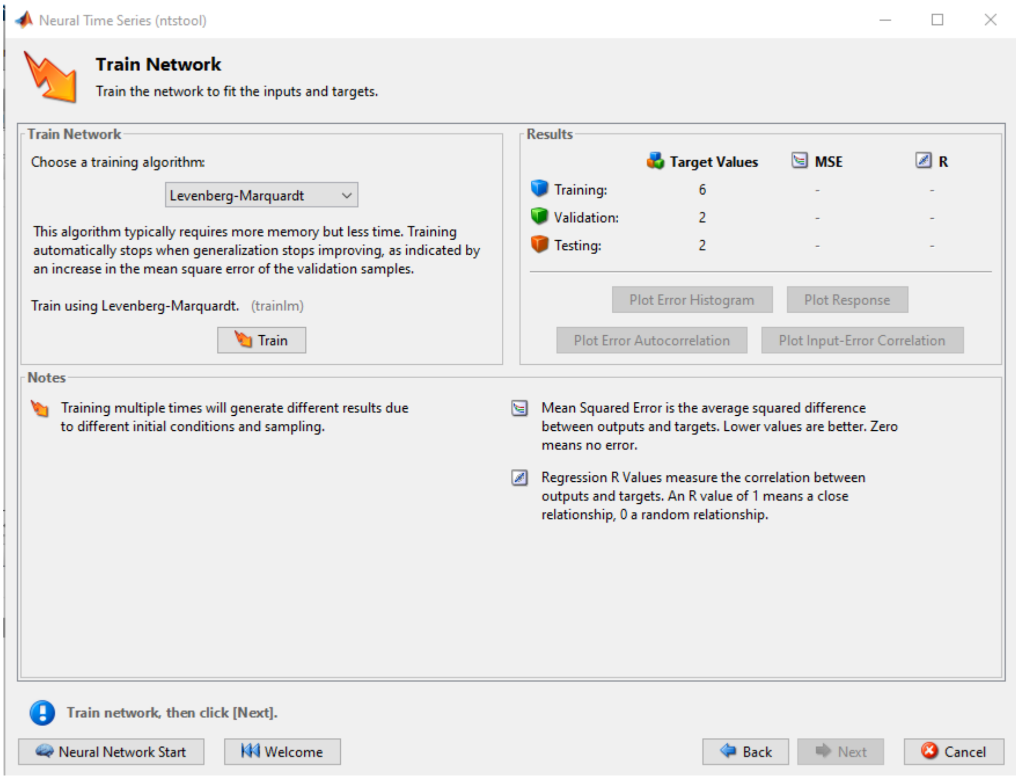

Figure 14. diagram showing training phase of ANN in MATLAB

4.6.4 Testing of the ANN network.

In this phase the network model is accessed based on its performance. Key performance indicators such as the Mean Square Error (MSE), Regression value (R). Other performance assessment tools such as Error Histogram, teaching performance, Error Autocorrelation, Training state, etc can be also used to give further evaluation on the performance of the model.

5 WASTE GENERATION MODELLING AND PREDICTION

5.1 Modelling of Municipal Solid Waste in selected cities of Ghana

To obtain the best network model different ANN structures with different hidden layers were tried during the training process and the best results for the modelling was for the network with 10 neurons. It was obtained in five iterations. The performance assessment tools were used to assess the model and it was observed that the network structure (4-10-1) had the least value for the Mean Square Error (MSE = 0.4) and the highest Regression Value (R = 0.9).

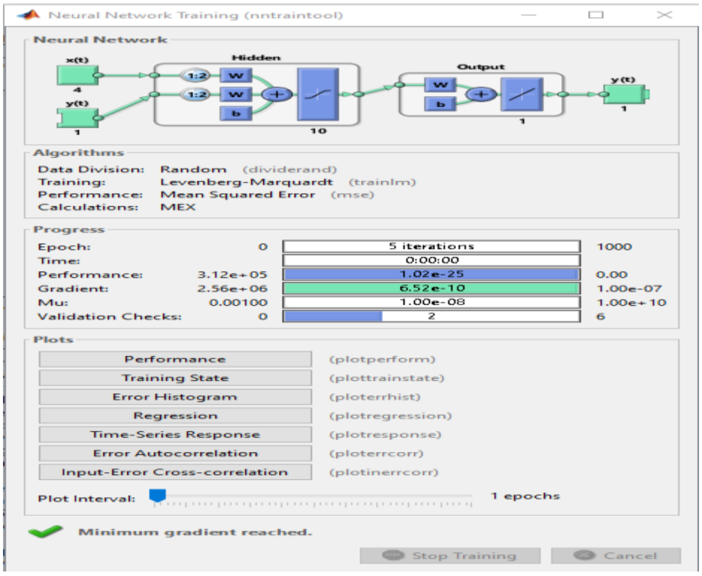

Figure 15. ANN structure for waste modelling in cities of Ghana.

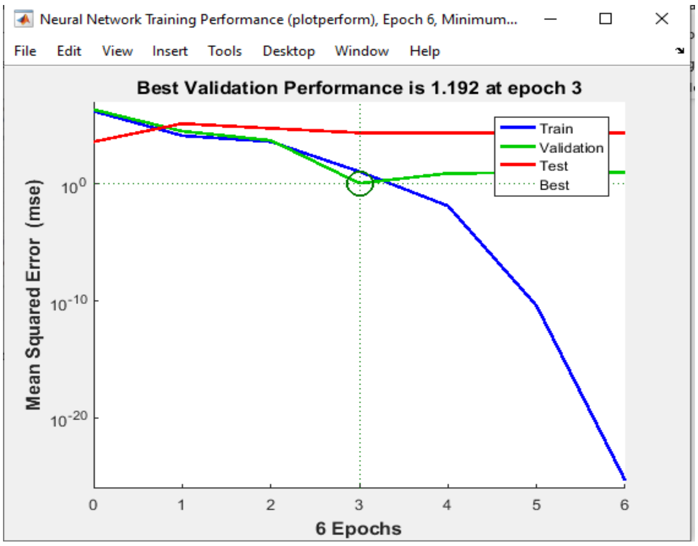

Figure 16. Teaching performance of the ANN

Figure 15 shows the teaching performance of the network it is observed that the training and validation errors reduces until the highlighted epoch. Also, there is no overfitting because the validation error does not increase before the epoch It is observed that during training of the network the error reduces.

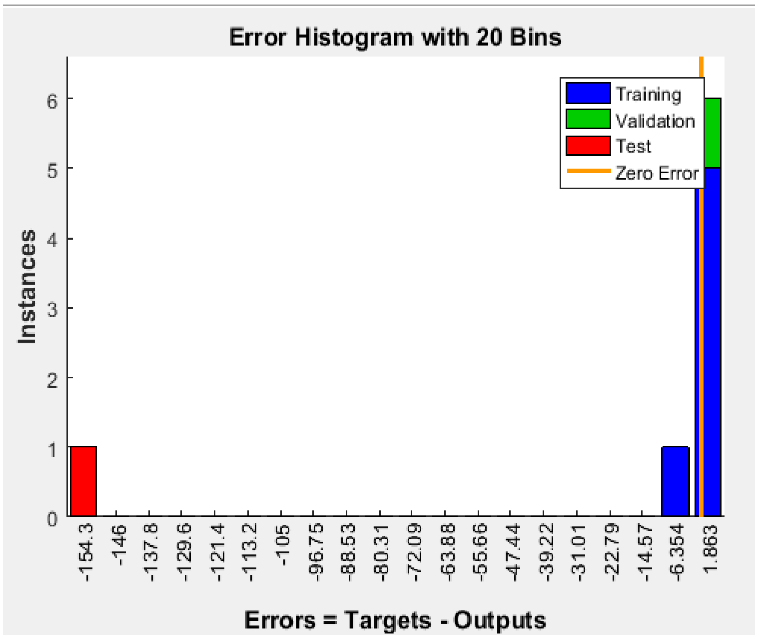

Figure 17. Error histogram of the network

The error histogram in figure 16 shows the error that occurred in training, validation and testing of the output and target. It further shows that the data fitting errors were spread within a good range between -5.354 and 1.863.

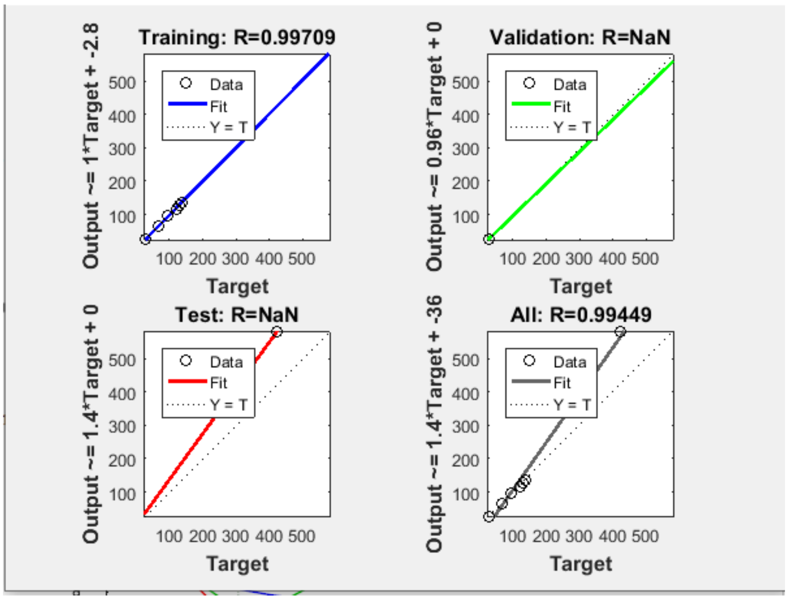

Figure 18. Regression values for the best model

Figure 15. shows the regression values of the network model. It can be observed that the R value for the training was R=0.997, for validation and testing R =Nan (Not a number) because there was not enough data for that. But the overall Regression value R=0.994 which shows a good fit of the network with the data.

5.2 Prediction of Municipal Solid Waste in selected cities of Ghana

The figure below shows the waste generation prediction for eight of the selected cities in Ghana. Time series forecasts uses previous data to predict the future. Time series forecasting is an easy approach which does not need much data. The predicted values from the modelling were plotted against the original data used (Table 1). It can be observed that the values of the actual and predicted were very close, meaning that the result is good. The prediction provided a good result based on the assessment of the performance indicators i.e., MSE= 0.43 and Regression values = 0.9.

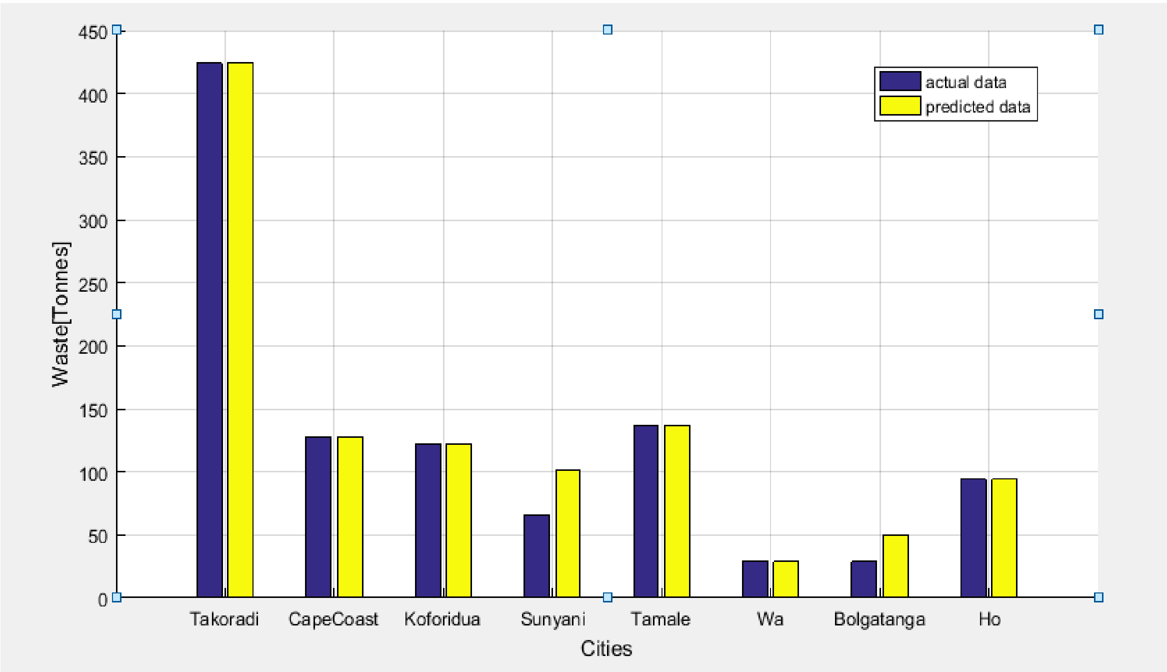

Figure 19. Waste Generation predictions for selected cities in Ghana

5.3 Modelling of total waste generated based on historical data

For modelling of the total waste generated on the historical data in Table 2, different ANN structures were also analysed to obtain the best results. The hidden layers were tried sequentially during the experiment the training process and the best results for the modelling was for the network with 8 neurons. It was obtained in 67 iterations. The performance assessment tools were used to assess the model and it was observed that the network structure (2-8-4) had the least value for the Mean Square Error (MSE) and the highest Regression Value (R).

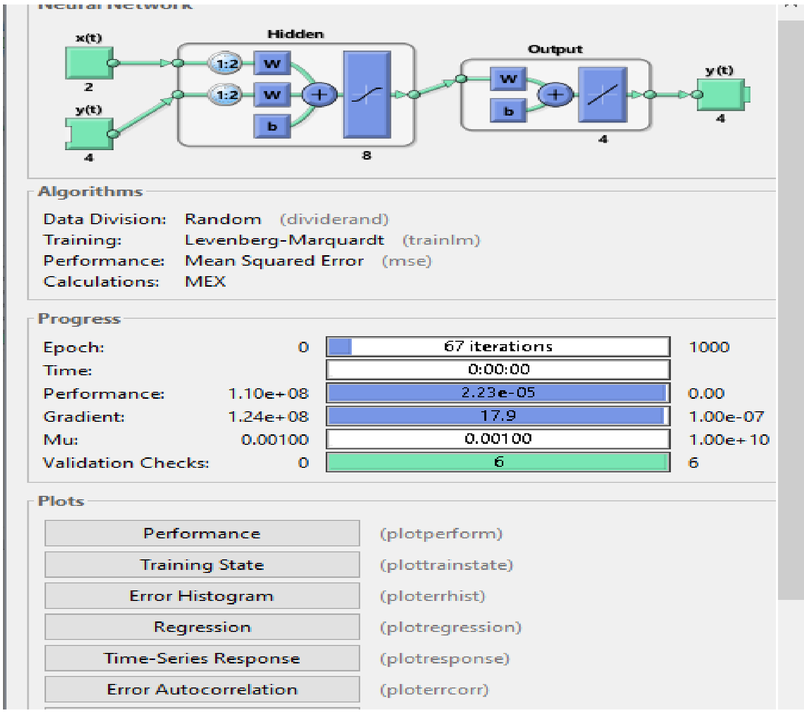

Figure 20. ANN structure for waste modelling in cities of Ghana on historical data

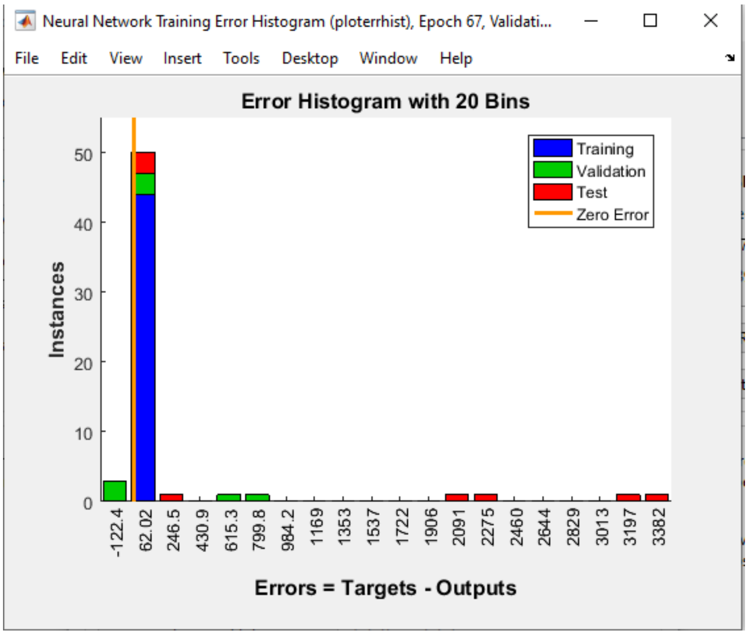

Figure 21. ANN structure showing error histogram of cities of Ghana on historical data.

The error histogram in Figure 18 shows the error that occurred in training, validation and testing of the output and target. It further shows that the data fitting errors were spread within a good range between -1.224 and 62.02.

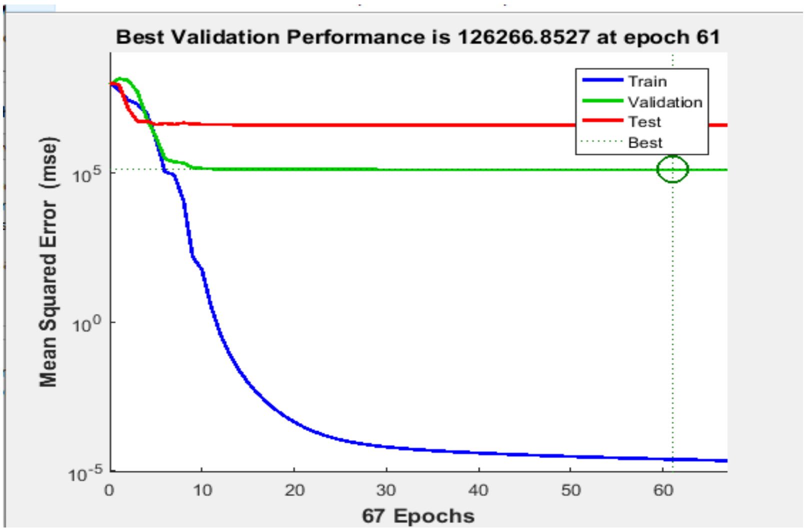

Figure 22. Teaching performance of the ANN

Figure 22 shows the teaching performance of the network the training and validation errors reduces until the highlighted epoch. Also, there is no overfitting because the validation error does not increase before the epoch. It is observed that the error reduces as the training progresses.

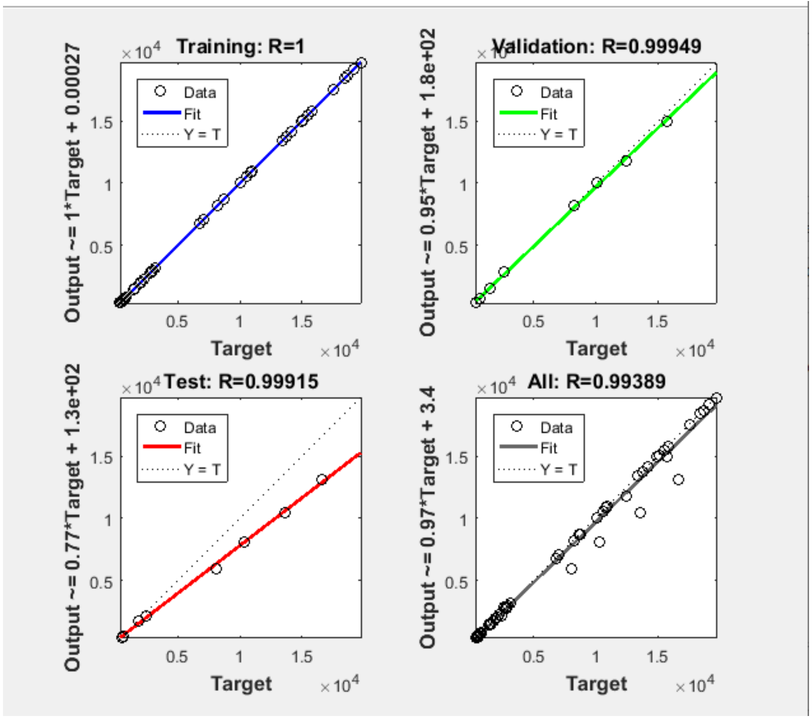

Figure 23. Regression values for the best model

Figure 23 above shows the regression values of the network model. It can be observed that the overall regression value R=0.994, for training R = 1, for validation R = 0.999 and testing R = 0.999. The overall Regression value R=0.994 shows a good fit of the network with the data.

5.4 Prediction of Municipal Solid Waste in Ghana from historic data

Figure 23 below shows the comparison between household waste and total waste generation from 2001 to 2015 in Ghana. It can be observed that the prediction was very similar to the actual data with just a small margin of error. Time series forecasting is an easy approach which does not need much data. The predicted values from the modelling were plotted against the original data used (Table 2). It can be observed that the values of the actual and predicted were very close, meaning that the result is good. The prediction provided a good result based on the assessment of the performance indicators i.e., MSE= 0.2 and Regression values = 0.993.

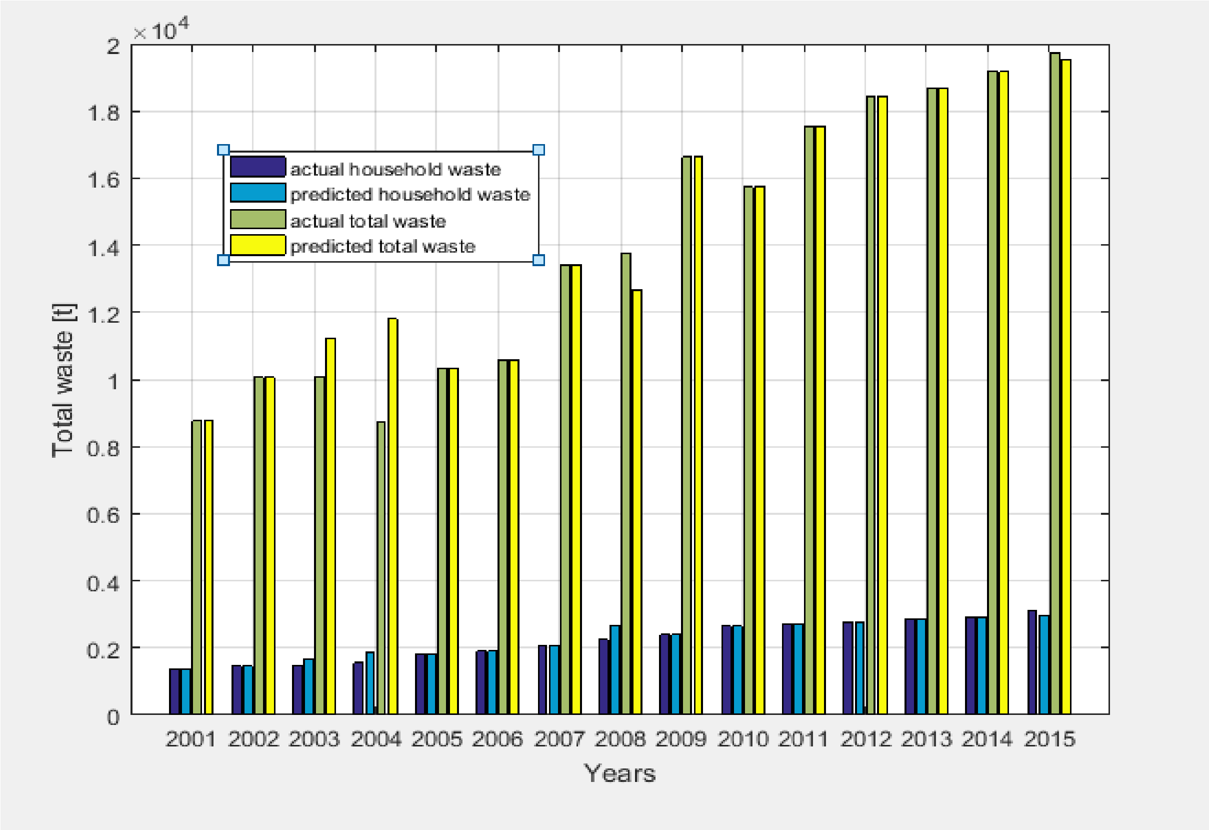

Figure 24. Total waste prediction based on historic data in Ghana.

5.5 Discussion of results

The study created two models. The first model predicted the total amount of generated waste based on socio- economic factors for some selected cities in Ghana. The second model predicted the total amount of waste in Ghana based on historical data collected from 1999-2015. It was observed that the selected socio- economic factors influenced the generation of waste. The second model also showed that increase in population over the years have direct effect on waste generation in Ghana. Two performance indicators namely Regression (R) and Mean Square Error (MSE) were used in accessing the performance of the models. The models with the best performance were selected as the best fit. The regression and MSE values of the best model (for waste generation based on socio-economic factors) were 0.994 and 0.43 respectively. The best model for (Total waste based on historical data) had regression and MSE values of 0.993 and 0.2 respectively.

Table 5. A comparison of other related studies on waste generation prediction

Area of Application	Case Study	Performance Metrics	References
Estimating weekly MSW output using ANN	Tehran	R = 0.837 and AARE = 0.044	[27]
Long-term prediction of solid waste generation using ANN	Mashhad, Iran	R = 0.86, MSE = 0.26, and MAPE = 0.046	[28]
Application of artificial neural networks for predicting the physical composition of municipal solid waste	Johannesburg, South Africa	R-values = 0.916,0.862, 0.834, and 0.826 for waste categories CE = 0.92, R = 0.96, WI = 0.98,	[29]
Prediction of Municipal Waste Generation in Poland Using Neural Network Modelling	Polish cities	R = 0.914, R = 0.989	[30]

In comparison with related studies by Younes, the best model had and MSE= 2.26 and R=0.914. in his study gross domestic product, employment and population were chosen as the input data [20]. A comparison between related studies is discussed below:

Adeleke in a similar study used ANN to predict the physical waste streams in Johannesburg using meteorological factors. Humidity, maximum/minimum temperature, wind speed was selected as the input data. The output variables forecasted were organic, plastics, paper, and textile waste. The best models had Regression values of 0.916, 0.886, 0.826 0.834 respectively [30].

Also, Ali Abdoli utilized ANN for a long term prediction of solid waste generation in the city Mashhad. Socio economic factors such as Population size, household income and temperature were identified as important in solid waste output. The best performing ANN model had Regression values of 0.86, MSE of 0.26 and Mean absolute percentage error (MAPE) of 0.046 respectively [28].

Nehal Elshaboury in his study also forecasted the generation of Municipal Solid waste (MSW) quantities using ANN-PSO. The best performing model had a Root Mean Square Error (RMSE) value of 11,342.74 and Mean Bias Error (MBE) value of 6548.55 [2].

CONCLUSION

Artificial Neural Networks which is a method of Big data processing can play a crucial role in Environmental Management. An accurate prediction of Municipal solid waste and total waste generated in Ghana is especially important to environmental management which influences the development of the nation. Increase in population, urbanisation, human lifestyles are one way or the other responsible for change in quantities of waste generation. In this study Artificial Neural network, a machine learning approach was applied on data sets obtained from Ghana. The results from the algorithms showed that indeed ANN can be an effective tool in predicting of waste generated in cities and the country.

The Artificial Neural Network model generated from the Neural Network toolbox in MATLAB showed a good predictive performance in the prediction of waste generated for some selected cities in Ghana and country. The estimation of waste generated is very crucial for environmental management because waste is one of the key factors that endanger the environment. Therefore, with the presented results the Government, private waste companies, environmental protection agencies can use ANN to develop models that would estimate the waste generation since these entities have substantial number of datasets that can be analysed using this Big data processing method.

If the input and output variables are correctly selected, machine learning approaches can be a powerful tool that can be used to make predictions on the generation of waste which can be particularly useful for planning and implementation of measures to protect the environment. For the limitations, the lack of enough data from cities in Ghana made the prediction a little biased. In addition, records of waste generation statistics should be made available so that better and more efficient models can be created to make better predictions. This study can also be used in the future with different machine learning algorithms to get better results. Optimisation algorithms can also be used to train the Artificial Neural Network for predictions.

BIBLOGRAPHY

[1] ESHUN, Isaac, 2013. Solid Waste Management: The Case of Komenda-Edina-Eguafo-Abirem (KEEA) Municipal in the Central Region of Ghana. Journal of Environmental and Earth Sciences. 2013. Vol. 3, no. 8, p. 1–11.

[2] ELSHABOURY, NEHAL, MOHAMMED ABDELKADER, ESLAM, AL-SAKKAF, ABOBAKR and ALFALAH, GHASAN, 2021, Predictive Analysis of Municipal Solid Waste Generation Using an Optimized Neural Network Model. Processes. 2021. Vol. 9, no. 11, p. 2045. DOI 10.3390/pr9112045. MDPI AG

[3] Lissah SY, Ayanore MA, Krugu JK, Aberese-Ako M, Ruiter RAC. Managing urban solid waste in Ghana: Perspectives and experiences of municipal waste company managers and supervisors in an urban municipality. PLoS One. 2021. 16(3) e0248392. doi: 10.1371/journal.pone.0248392. PMID: 33705483; PMCID: PMC7951920.

[4] GLASGOW, T H E, 1824. the Glasgow. . 1824. No. V.

[5] ISO 14001:2015(EN) Environmental managements systems- Requirements with guidance for use [online], 2015 [cit. 2023-03-24]. Available from: https://www.iso.org/obp/ui/#iso:std:iso:14001:ed-3:v1:en:en/

[6] Goal 13: Take urgent action to combat climate change and its impacts [online], 2023 [Cit.2023-04-15]. Available from: https://www.un.org/sustainabledevelopment/climate-change/

[7] Basel Convention on the Control of Transboundary Movements of Hazardous wastes and their Disposal [online], 2023 [cit. 2023-04-04]. Available from: http://www.basel.int/TheConvention/Overview/TextoftheConvention/tabid/1275/Default.aspx

[8] Waste Management: An important tool to ensure environmental sustainability [online], 2022. [cit. 2023-03-17] Available from: https://centurionlg.com/2021/11/22/waste-management-an-important-tool-to-ensure-environmental-sustainability/

[9] Big Data and Environment- Green data: Can statistics help the environment? [online], 2023 [cit. 2023-03-24] Available from: https://www.iberdrola.com/sustainability/big-data-and-environment.

[10] Dubey, Rameshwar & Gunasekaran, Angappa & Childe, S. & Papadopoulos, Thanos & Luo, Simon & Fosso Wamba, Samuel & Roubaud, David. (2017). Can Big Data and Predictive Analytics Improve Social and Environmental Sustainability. Technological

Forecasting and Social Change. [online], 2017 [cited 2022-05-16]. Available from: https://www.researchgate.net/publication/318446913

[11] Big Data and Cognitive Computing [online], 2022 [cit. 2022-03-24]. Available from: https://www.mdpi.com/journal/bdcc

[12] Reihner H. Hariri, Erik M. Fredericks, and Kate M. Bowers [online], 2019 [cit. 2022-05-16]. Available from https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0206-3#citeas

[13] Big data Analysis Techniques[online], 2022 [cit. 2022-03-25]. Available from https://www.getsmarter.com/blog/career-advice/big-data-analysis-techniques/

[14] How do data fusion technologies work? [online], 2022. [cit. 2022-05-18] Available from: https://sdi.ai/blog/how-do-data-fusion-technologies-work/

[15] What is Data Processing [online], 2022. [cit. 2022-05-17] Available from: https://www.talend.com/resources/what-is-data-processing/

[16] KANNANGARA, Miyuru, DUA, Rahul, AHMADI, Leila and BENSEBAA, Farid. Modelling and prediction of regional municipal solid waste generation and diversion in Canada using machine learning approaches. Waste Managemen*t* [online], 2018. [cit. 2022-03-13] Available from: https://doi.org/10.1016/j.wasman.2017.11.057

[17] Oribe, Iraia & Kamara-Esteban, Oihane & Martin, Cristina & Macarulla, Ana & Alonso-Vicario, Ainhoa. Identification of influencing municipal characteristics regarding household waste generation and their forecasting ability in Biscay. Waste Management.

[18] Daskalopoulos, E., Badr, O., & Probert, S.D. 1998. An integrated approach to municipal solid waste management. Resources Conservation and Recycling, [online] 2009. [cit. 2022-05-12] Available from: https://doi.org/10.1016/S0921-3449(98)00031-7

[19] ABBASI, M., ABDULI, M.A., OMIDVAR, B. and BAGHVAND, A. Results uncertainty of support vector machine and hybrid of wavelet transform-support vector machine models for solid waste generation forecasting. Environmental Progress & Sustainable Energy. 2013. Vol. 33, no. 1p. 220–228. DOI 10.1002/ep.11747.

[20] Ordonez-Ponce, Eduardo & Samarasinghe, Sandhya & Torgerson, L. Artificial neural networks for assessing waste generation factors and forecasting waste generation: A case study of Chile. Journal of Solid Waste Technology and Management. 2006. 32. 167-184.

[21] Prediction of Municipal Solid Waste Generation by Use of Artificial Neural Network: A Case Study of Mashhad [online], 2007 [cit. 2022-05-12] Available from: https://ijer.ut.ac.ir/article_170.html

[22] YANG, Chao, WU, Guofeng, DING, Kai, SHI, Tiezhu, LI, Qingquan and WANG, Jinliang. Improving land use/land cover classification by integrating pixel unmixing and decision tree methods. Remote Sensing. 2017. Vol. 9, no. 12p. 1222. DOI 10.3390/rs9121222.

[23] Firat, Turan, Yurdusev, Comparative analysis of neural network techniques for predicting water consumption time series. Journal of Hydrology Volume 384, Issues 1–2, 15 April 2010, Pages 46-51

[24] Kodwo Miezah, Kwasi Obiri-Danso, Zsófia Kádár, Bernard Fei-Baffoe, Moses Y. Mensah, Municipal solid waste characterization and quantification as a measure towards effective waste management in Ghana, Waste Management, September 2015, Volume 46,2015, Pages 15-27, ISSN 0956-053X,

[25] Ghana statistical Service, Population statistics [online] 2023 [cit. 2023-04-14] Available from: https://www2.statsghana.gov.gh/pop_stats.html

[26] Ali, S.A., Ahmad, A. Forecasting MSW generation using artificial neural network time series model: a study from metropolitan city. SN Appl. [online], 2019 [cit. 2023-03-26]. Available from https://doi.org/10.1007/s42452-019-1382-7

[27] Nouri Rouh Elah, Abdoli M.A, Farokhina Askan, Ghaemi Alalleh. Solid waste Generation by hybrid of artificial neural network and wavelet transform. Journal of environmental studies. [online], 2009 [cit. 2023-03-25]. Available from https://www.sid.ir/en/journal/ViewPaper.aspx?id=140867

[28] Ali Abdoli, M.; Falah Nezhad, M.; Salehi Sede, R.; Behboudian, S. Longterm forecasting of solid waste generation by the artificial neural networks. Environ. Prog. Sustain. Energy [online], 2011 [cit. 2023-05-12]. Available from: http://aiche.onlinelibrary.wiley.com/doi/10.1002/ep.10591

[29] Adeleke, O., Akinlabi, S. A., Jen, T.-C., & Dunmade, I. Application of artificial neural networks for predicting the physical composition of municipal solid waste: An assessment of the impact of seasonal variation. Waste Management & Research [online], 2021 [cit. 2023-04-13]. Available from: https://doi.org/10.1177/0734242X21991642

[30] KULISZ, Monika a Justyna KUJAWSKA. Prediction of Municipal Waste Generation

in Poland Using Neural Network Modeling. Sustainability [online], 2020 [cit. 2023-03-15] Available from: http://doi:10.3390/su122310088

LIST OF ABBREVIATIONS

MSWM	Municipal Solid Waste Management
ANN	Artificial Neural Networks
DBSCAN	Density based spatial clustering of applications with noise
MSE	Mean Square Error
SQL	Structured Query Language
NoSQL	Non-relational Structured Query Language
GDP	Gross domestic product
MSW	Municipal solid waste
SVM	Support vector mechanics
ML	Machine learning
DT	Decision trees
NARX	Nonlinear auto regressive

LIST OF FIGURES

LIST OF TABLES

APPENDICES

APPENDIX PI: CD -ROM