NEAR EAST
UNIVERSITYINSTITUTEOFGRADUA
TESTUDIES
DEPARTMENTOFBIOMEDICALENGINEERING

MACHINELEARNINGAPPROACHESFORPREDICTINGMALARIAANDMONKEYPOXDISEASES

Ph.D.THESIS

BASILBARTHDUWA

Nicosia
December, 2023

NEAR EAST
UNIVERSITYINSTITUTEOFGRADUA

DBASILBARTH
DUWA

MACHINELEARNINGAPPROACHESFORPREDICTINGM
ALARIA AND MONKEYPOXDISEASES

PHDDISS
ERTATION

2023

DEPARTMENTOFBIOMEDICALENGINEERING

MACHINELEARNINGAPPROACHESFORPREDICTINGMALARIAANDMONKEYPOXDISEASES

Ph.D.THESIS

BASILBARTHDUWA

Supervisor

Assoc.Prof.Dr.DilberUzunOzsahin
Co-Supervisor
Assist.Prof.Dr.BernaUzun

NicosiaDecember,2023

Approval

We certify that we have read the thesis submitted by **BASIL BARTH DUWA** titled **"MACHINE LEARNING APPROACHES FOR PREDICTING MALARIA AND MONKEYPOX DISEASES."** and that in our combined opinion it is fully adequate, in scope, and quality, as a Dissertation for the degree award of Doctor of Philosophy in Biomedical Engineering.

Examining Committee	Name-Surname	Signature
Head of the Committee:	Assoc. Prof. (Dr.) Süleyman Aşır	…………………………
Committee Member*:	Prof. (Dr.) Kerem Cankocak	…………………………
Committee Member*:	Prof. (Dr.) Isa Dumanoğlu	…………………………
Committee Member*:	Assoc. Prof. (Dr.) Doğa Kavaz	…………………………
Supervisor:	Prof. (Dr.) Dilber Uzun Ozsahin	…………………………

Approved by the Head of the Department

15 /01/2024

Assoc. Prof. Dr. Süleyman Aşır

…………………………………

Head of Department

Approved the Institute of Graduate Studies

…. /…. /2023

Prof. Dr. Kemal Hüsnü Can Başer

…………………………………

Head of the Institute

Declaration

I hereby declare that all information, documents, analysis and results in this thesishavebeencollectedandpresentedaccordingtotheacademicrulesandethicalguidelinesofInstituteofGraduateStudies,NearEastUniversity.Ialsodeclarethat asrequiredbytheserulesandconduct,Ihavefullycitedandreferenced informationanddatathatarenotoriginaltothisstudy.

BasilBarthDuwa
05…/…01…/…2024
Day/Month/Year

Acknowledgement

As I approach this momentous milestone, I am overwhelmed with tremendous gratitude foreveryone who has played a role in this journey. This accomplishment wouldn't have beenfeasiblewithoutthe help,motivation,forwhichIamgrateful.

Tomyloving father Mr.BarthDuwa,yourpersistentconvictioninmyabilityandunwaveringsupport,bothmentallyandacademically,wereadrivingforceunderlyingmyaccomplishments.

TomyLatecherishedmotherMrs.ChristyDuwa,although yourpresenceisnolongerwithus,the memory of you and the sacrifices you made during my studies will live on in my heart.youraffectionandsupportwereabeaconofhopeinmyacademiccareer.

I cannot forget Assoc. Prof. Dr. Dilber Uzun Ozsahin my supervisor for not only supportingme academically but also for her unwavering support and encouragement throughout thischallengingjourney.Yourguidancehasbeeninvaluable.

Words cannot express my genuine appreciation to my co-supervisor, Assist. Prof. Dr. BernaUzun for her enthusiasm and passion shown on me, specially. She has been a guiding force inmy academic pursuits, from my master's level to this defining moment. Your professionalmentorship and the familial atmosphere you provided have enriched my academic experienceimmensely.Ihonestlydonottakethatforgranted.

I extend my appreciation to Assoc. Prof. Dr. Suleyman Asir, my Head of Department, foractively participating in my academic growth and fostering an environment conducive tolearning and research. I also appreciate the Departmental, faculty and school board membersfortheireffortsinmakingthisinstitutionstand-outas thebest.

I specially thank Dr. Ilker Ozsahin for the courses I was privileged to be tutored by him, forthecorrectionsIfindworthy.

I would like to express my gratitude to my numerous colleagues and team members at NearEast University, Operational Research Centre. Your camaraderie and collaborative spirit havecontributedtomyacademicandpersonalgrowth.

Thisjourneyhasbeenatestamenttothepowerofsupport,mentorship,andperseverance.Iamimmenselygratefultoeachoneofyoufor yourcontributions,encouragement,andbeliefinmypotential.

Abstract

“MACHINELEARNINGAPPROACHES FORPREDICTINGMALARIAANDMONKEYPOXDISEASES.”

BASILBARTHDUWA

PhD,DepartmentofBiomedicalEngineeringD
ecember,2023.

According to the World Health Organisation (WHO), infectious diseases like malaria andmonkeypox have resulted in catastrophes. Malaria is typically spread by mosquito bites frominfected individuals, while monkeypox is developing as a zoonotic disease with feasibletransmission to humans. Malaria and monkeypox are transmissible illnesses caused by bothparasites and viruses, respectively. Utilizing modern techniques for these diseases’ detection,diagnosis,prevention,andprognosishasbecome essential.

This study uses a variety of clinical and demographic characteristics to develop machine-learningmodelsforthepredictionoftheseinfectiousdiseases.Adatasetof20variables,whichinclude stemperature,theresultsofdiagnostictests,bloodcellcounts,andotherhematologicalfactors, is examinedformalaria. While,for monkeypox, a dataset thatincludes SystemicIllness,Rectal Pain,Sore Throat, Penile Edema,Oral Lesions,Solitary Lesion,SwollenTonsils, HIV Infection, Sexually Transmitted Infection and Monkeypox. Machine learningmodels such as Logistic Regression, Naive Bayes, Random Forest, and Decision Treewereadopted in the prediction, while the evaluation of performance was done using Accuracy,Precision, Recall and F1-score. The total number of malaria cases in the dataset is 8,250, ofwhich 4,953 are positive and 3,297 are negative. While, there are 25,000 instances in themonkeypoxdataset,16,160ofwhicharepositiveand8,840ofwhicharenegative.Toincreasemod elresilience,datapreparationmethodssuchasone-hotencoding,scaling,anddatasplittingareused.

The result shows the greatest accuracy of 91% and balanced precision, recall, and F1 scores of0.92, 0.91, and 0.91 in the case of malaria prediction, Linear Regression showed superiorpredictiveability.NaiveBayesmaintainedaprecision,recall,andF1scoreat0.90, 0.89, and 0.89whileachievingalittleloweraccuracyof89%.Withanaccuracyof93%,precision,recall,andF1s coresof0.90,0.89,and0.89,RandomForestsurpassedothermodels,demonstrating itsdependability.DecisionTreedemonstratedsuperiorperformancewithprecision,recall,andF1scores of0.90,0.89,and0.89,respectively.Itsaccuracywas 88%.

Exceptional performance for monkeypox prediction was shown by the evaluation results ofDecision Tree Classifier, Random Forest Classifier, Naive_bayes, and Logistic Regression.Eachmodelreceivedaperfectaccuracyratingof1.0,meaningthateverytestsetexample wasproperly identified. Precision, recall, and F1-score, all of which had values of 1.0 for bothclasses (0 and 1), were considered measures of robustness. The F1-score emphasised greatmemoryincapturingallactualcasesofmonkeypoxandgoodaccuracyincorrectlyidentifyingcases of monkeypox, demonstrating a fair balance between precision and recall. The findingsof the study demonstrate how successful these categorization algorithms predict and classifyinstancesofmalariaandmonkeypox,providingimportantinformationfordiseasemanagementandpublichealthinitiatives.

Keywords: Infectiousdiseases,malaria,monkeypox,machinelearning,predictivemodelling.

ÖZET

DünyaSağlıkÖrgütü’ne(WHO)göresıtmavemaymunçiçeğigibibulaşıcıhastalıklarfelaketlerlesonuçlanmıştır.Sıtmatipikolarakenfektekişilerdensivrisinekısırıklarıylayayılırken, maymun çiçeği insanlara bulaşması mümkün olan zoonotik bir hastalık olarakgelişiyor. Sıtma ve maymun çiçeği sırasıyla parazitlerin ve virüslerin neden olduğu bulaşıcıhastalıklardır. Bu hastalıkların tespiti, tanısı, önlenmesi ve prognozunda modern tekniklerinkullanılmasızorunluhalegelmiştir.

Bu çalışmada, bu bulaşıcı hastalıkların tahmini için makine öğrenimi modelleri geliştirmekamacıyla çeşitli klinik ve demografik özellikleri içeren verilerden yararlanılmıştır. Kişilerinateşölçümleritestlerininsonuçları,kanhücresisayımlarıvediğerhematolojikfaktörleriiçeren20 değişkenden oluşan bir veri seti sıtma için incelenmiştir. Maymun çiçeği hastalığı için isesistemikhastalık,rektal ağrı ve ağızlezyonlarını içeren değişkenleri içeren birveri setikullanılmıştır.

Tahminde Lojistik Regresyon, Naive Bayes, Rastgele Orman ve Karar Ağacı gibi makineöğrenimi modelleri uygulanırken, performans değerlendirmesi doğruluk, kesinlik, duyarlılıklıkve F1 puanı kullanılarak yapılmıştır. Veri setindeki toplam sıtma vakası sayısı 8.250 olupbunların4.953’ü pozitif,3.297’siisenegatiftir.Maymun çiçeğiveri setindeiseverilerin16.160’ı pozitif, 8.840’ı negatif olmak üzere 25.000 örnek içermektedir. Model esnekliğiniartırmak için tek-etkin kodlama, ölçeklendirme ve veri bölme gibi veri hazırlama yöntemlerikullanılmıştır.

Sonuçolarak,sıtmahastalığıtahmininde%91’likenyüksek doğrulukve0,92,0,91ve0,91’likkesinlik, duyarlılık ve F1 değerleri elde edilmiştir; kullanılan modeller arasında DoğrusalRegresyon modeli bu konuda üstün tahmin yeteneği göstermiştir. Naive Bayes modeli ilekesinlik,duyarlılıkveF1değerleri0,90,0,89ve0,89olarakeldeedilirkenbumodelile %89’luk biraz daha düşük bir doğruluk elde etmiştir. %93’lük doğruluk, 0,90, 0,89 ve 0,89’lukkesinlik, duyarlılık ve F1 puanlarıyla Rassal Orman modeli diğer modelleri geride bırakaraksıtma hastalığı tahmininde güvenilir bir model olduğunu göstermiştir. Karar Ağacı modeli desırasıyla 0,90, 0,89 ve 0,89’luk kesinlik, duyarlılık ve F1 değerleri ileüstün performansgöstermiştir.Bu modelileelde edilendoğrulukise%88olarakeldeedilmiştir.

MaymunçiçeğihastalığıtahminiiçinbaşvurulanKararAğacıSınıflandırıcısı,RastgeleOrmanSınıflandırıcısı,Naive_bayesveLojistikRegresyonunmodellerininüstünperformansgösterdiği gözlenmiştir. Belirtilen tüm modeller ile, 1,0’lik mükemmel bir doğruluk derecesielde edilmiştir; bu, her test seti örneğinin doğru şekilde tanımlandığı anlamına gelmektedir.Tanımlanan her iki sınıf için (pozitif vaka veya değil) güvenilirlik ölçüleri olarak elde edilenkesinlik, duyarlılık ve F1 puanları 1,0 dolarak elde edilmiştir. F1 değeri, maymun çiçeğivakalarını yakalamada ve doğru bir şekilde tanımlamada modellerin mükemmel bir hafızagösterdiğinin göstergesidir. Kesinlik değerinin 1,0 olarak elde edilmesi iyi bir doğruluk eldeedildiğinin göstergesidir ve kesinlik ile duyarlılık arasında adil bir denge elde edilmiştir.Çalışmanınbulguları,busınıflandırmaalgoritmalarınınsıtmavemaymunçiçeğiörneklerininekadar başarılı bir şekilde tahmin edip sınıflandırdığını, hastalıkyönetimi ve halk sağlığıgirişimleriiçinönemlibilgilersağladığınıgöstermektedir.

AnahtarKelimeler: Bulaşıcıhastalıklar,sıtma,maymunçiçeği,makineöğrenimi,tahminedayalımodelleme.

Table ofContents

ListofTables

ListofFigures

Abbreviations

Abbreviations	Meaning
MP	Malaria
MPOX	Monkeypox
WHO	WorldHealthOrganization
MPXV	MonkeyPoxvirus.
HIV	HumanImmune deficiencyvirus
STI	Sexually-transmittedinfection
NSAIDs	Non-SteroidalAnti-inflammatoryMedicines
RDC	DemocraticRepublicofCongo
ML	Machinelearning
RDTs	Rapiddiagnostictests
PCR	Polymerasechainreaction
ITNs	Insecticide-treatedbednets
IRS	Indoorresidualspraying
ACTs	Artemisinin
MCDM	Multi-criteriadecision-making
PROMETHEE	PreferenceRankingOrganisationMethodforEnrichmentEvaluation
TOPSIS	FuzzyTechniqueforOrderofPreferencebySimilaritytoIdealSolution
CT	Computedtomography
SEL	EnsembleLearning
DL	DeepLearning
CNN	Convolutionalneuralnetwork
Nmi	Non-malarialinfection
SM	Severemalaria
UM	Uncomplicatedmalaria
ANN	Artificialneuralnetwork
RBC	Redbloodcell
AI	Artificialintelligence
MRI	MagneticResonanceimaging
HER	AnalysisofElectronicHealthRecords

MSE	Meansquared error
MAE	Meanabsolute error
R2	R-squared
DTs	DecisionTrees
TP	TruePositives
TN	TrueNegatives
FP	FalsePositives
FN	FalseNegatives
ROC	ReceiverOperatingCharacteristic
EDA	Exploratorydataanalysis
PCA	Principalcomponentanalysis
CCA	Canonicalcorrelationanalysis
SEM	Modellingofstructuralequations
LR_MOD	LogisticRegressionclassifier
GB_MOD	GaussianNBclassifier
RF_MOD	RandomForestClassifier
DT_MOD	DecisionTreeClassifier
URTIs	Upperrespiratorytractinfections

CHAPTER1

Introduction

1.1 BackgroundofStudy

Infectiousailmentscontinuetoposeasignificantthreattothepublichealth,especiallyinruralareas that have restricted access to medical services. Malaria (MP) and monkeypox (mpox),among many such illnesses, continue to wreak havoc on populations, primarily in regard tohealth and financial loss. The importance of precise and timely disease prediction cannot beemphasized, as it is the foundation of efficient disease control and prevention. Conventionaldiagnostic procedures, on the other hand, frequently fall short of giving accurate outcomes,especiallyinpoorregionswheresuchillnessestendtobe mostsevere[1].

In2022,theWorldHealthOrganization(WHO)reportedthatanoutbreakofmonkeypoxaroseabruptly and quickly spreading over Europe, the Americas, and eventually all six WHOregions, with over 100 countries reporting approximately 87 thousand instances and 112mortality. The global epidemic has mostly (but not only) afflicted gay, bisexual, and othermales who engage in intercourse between men, and it propagates between individuals viainterpersonalcontact[2].

Monkeypox,commonlyreferredtoasmpox,isazoonoticillnesscarriedbytheinfectioncalledmonkey virus, which belongs to a member of the Orthopoxvirus genus (MPXV).Numerousspecies of wild mammals make up the reservoir. Only Cameroon has reported all MPXVgenetic clades,namelyCladeIinCentralAfricaaswellasClade IIin WestAfrica.

Most instances of human infection have been linked to human migration, transportation byinfected people, and contact with foreign sick wildlife [3]. Human cases have mostly beenlinked to rural, forests in various central and west African countries. Since 2014, there havebeenmorecasesreportedeach yearinAfricathantherewereinreportsfromthe40 yearspriortothatinnearlyallofthecontinent[4].Theresurgenceofmpoxcouldbebroughtonbya mixofenvironmentalandsocioeconomicfactors,animalorhumanmigration.Theendofsystematic smallpox immunisation since the disease was eradicated in 1980, advancements indiseasedetectionandassessment,andgeneticalterationsinthevirus[5].

1.1.1 SymptomsofMonkeyPox

The symptoms linked with a MonkeyPox are presented in Table 1:1. Fever is among the mostprevalentsymptoms,beingreportedby86%ofpeople,demonstratingitsconsiderablepresence. The rash comesin secondat84%,indicatinga high prevalenceof cutaneoussymptoms. Headaches are reported by 63% of people, emphasizing the disorder’s neurologicsignificance. Fatigue, reported by 47% of respondents, emphasizes the systemic character ofthe symptoms. Lymphatic nodes swell (36%), as well as muscular pains (32%), indicateimmunologicalandmusculoskeletalresponses.Chills(27%)andperspiration(21%)arenaturalreactionstothesituation.Asorethroat,asreportedby15%,showstherespiratorysystemmaybe involved. This information is useful for understanding symptoms and assisting healthcarepractitionersindiagnosisandmanagementtechniques[6].

Table1:1

SymptomsofMonkeyPox[6]

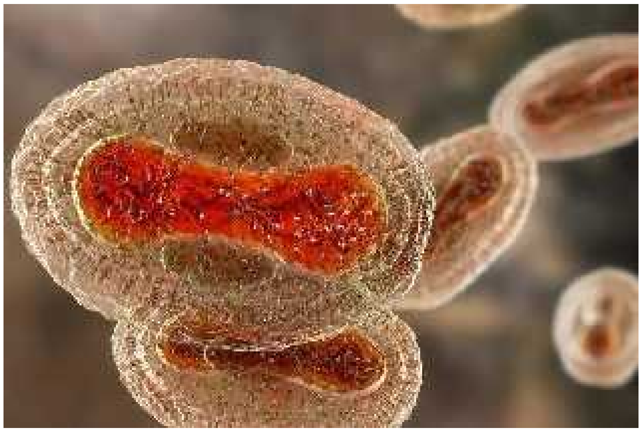
symptoms	**Percentage(%)**
Fever	86
Rash	84
Headache	63
Fatique	47
Swellingoflymphnodes	36
MuscleAches	32
Chills	27
Sweating	21
Sorethroat	15

Thevarietysymptomsandindicatorscanbecausedbymumps.Whilecertainindividualsexperiencelessseveresymptoms,othersmayexperiencemoreseverediseasesandrequire medicalattention.Pregnantwomen,youngchildren,andthosewithimpairedimmunesystems,such as those with HIV infection, are often more likely to experience more severe symptoms[7].

Figure 1:1 shows the 3D illustration of monkeypox virus under a microscope. A skin eruptionthat could persist for up to four weeks is one of the common signs of mpox. Fever, headache,muscle aches, back discomfort, low energy, and enlarged glands (lymph nodes) may appearfirst or be followed by this. The face, groyne, genital, and anal regions can all be affected bythe rash, which has a blister-like or painful appearance [8]. These lesions can also develop onthe eyes, mouth, throat, anus, rectum, or vagina. From one to several thousand sores may bepresent.Proctitis,aswellingoftherectumthatcanbeextremelypainful,andgenitalinflammation thatcan makeitdifficulttourinate are both conditions thatsome peopleexperience[9].

Figure1:1

Monkeypoxvirus,3Dillustration[10]



1.1.2 TransmissionofMonkeyPox

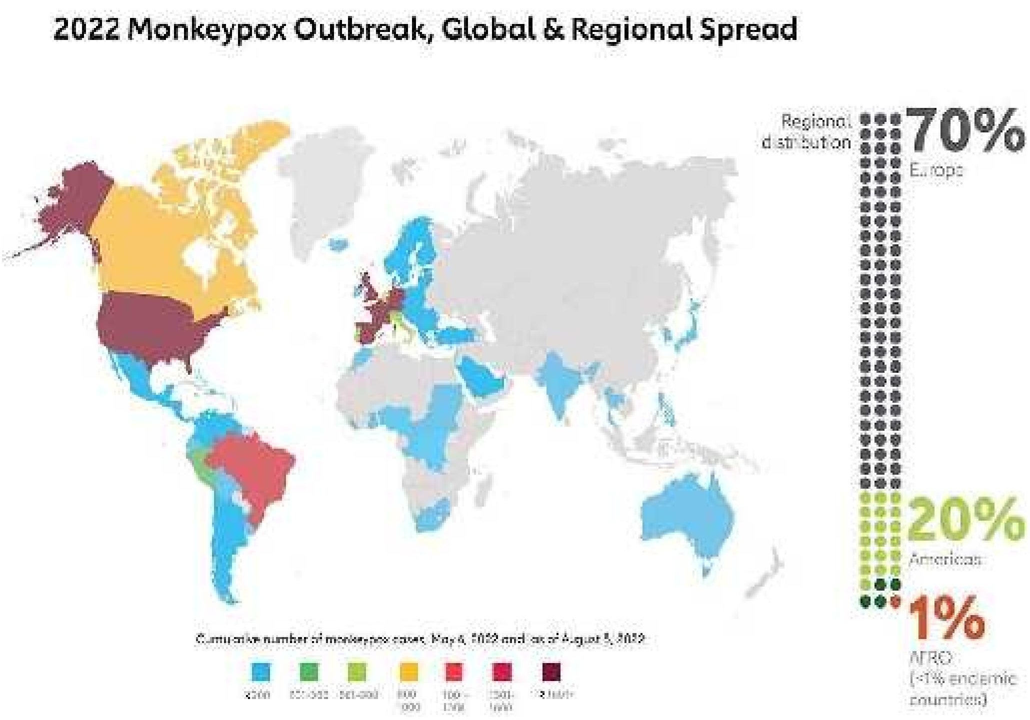
Personalcontactwithinfectedanimals,especiallyrodentsandmonkeys,aswellascontactwiththeir body fluids or lesions are the main ways that monkeypox spreads from infected animalsto human beings. Having close contact with the blood, body fluids, or wounds of an infectedindividual can result in transmission between people [11]. During coughing or sneezing,airbornedropletscanpotentiallybeamethodofinfection.Althoughitistheoreticallyfeasible,straight transmission from animal to human is more effective than transmission betweenhumans.Preventingcontactwithanimalsthatmaycarrythevirus,adoptingproperhand hygiene,utilizingprotectivegearwhentakingcareofinfectedpeople,andinstitutingconfinement and quarantine procedures during epidemics are all useful preventative strategies[12].

1.1.3 MonkeyPoxEpidemiology:

In past epidemics, the Republic of Congo had six instances, which represented the longestcontinuous sequence of transmission between individuals or the most consecutive human-to-humancases.Now,everythingappearstobe extremelydifferentasshown inFigure1:2. Nearlyonethousandinstancesofmonkeypoxarerecordedgloballyapproximatelydaily,andthenumber ofinstancesisstillrising[14].Withjust overonepercentofthecasesreported atthis time coming from epidemic nations, this epidemic is also distinct in that the data nowaccessible showsitismostlyexpandinginEuropeandthe Americas[15].

Figure1:2

Global Outbreak[13].



Thebulkofnewinfectionscontinuestoberestrictedtofivedifferentcountries,despitethefactthat the total number of nations confirming incidents has increased from a small to highnumbers[16]. According to the WHO, the predicted incidence is 1.8 in Spain’s metropolitan areas, 1.6 in theUnited Kingdom, and 1.4 in Portugal. In Europe, the majority of those with monkeypox areimmunosuppressed; 38–40% of those with positive tests are also HIV+. In Europe, 94 percentof infections are not considered severe, while only 5 to 6 percent of infected people requirehospitalizationasinFigure1:2[17].

1.1.4 MonkeyPoxDiagnosis:

Monkeypox is normally diagnosed using a mix of clinical assessment, laboratory tests, andanalysisoftheindividual’spastmedicalconditionsandsigns[18].Thefollowingarethemainindicatorsfordiagnosingmonkeypox:

i. ***The Clinical Assessment:*** Medical professionals evaluate the patient’s symptomsandtheresultsoftheirmedicalexam.Theusualsignsofmonkeypox includefever,rash, enlarged lymph nodes, and pustular lesions on the skin. When monkeypox isinitiallysuspected,acomprehensiveclinicalassessmentisveryimportant[19].

ii. ***Patient History:*** The medical records of the individual in question, specificallysignificant recent contact with animals (especially rodents or monkeys) or peopleexhibitingcomparablesymptoms,areessentialinformation forthediagnosis [20].

iii. ***Laboratory Tests:*** Testing in laboratories, such as the ones below, are frequentlyusedtovalidateadiagnosis.

   - PCR (Polymerase Chain Reaction) testing can identify the genetic makeup of themonkeypoxvirusinbloodstream,lesionsoftheskin, orother medicalspecimens.
   - ViralCulture:Viralisolationinalabsettingusingviralcultureisanothertechniquefordiagnosingmonkeypox.
   - Serological Examinations: Samples of blood could be examined for the detectionof certain antibodies directed targeting antigens of the monkeypox virus. Antibodylevelsthatgraduallygrowmaybe a signofongoinginfection[21].

iv. ***DifferentialDiagnosis:***Thesymptomsofmonkeypoxcanbemistakenforthoseofotherviraldiseases,suchaschickenpox,smallpox(whichisgenerallyexterminated but can recur sometimes), and a number of bacterial skin infections. It's crucial to rule out any further potential explanations of symptoms that are similar[22].

1.1.5 Prevention of Monkey Pox:

Primarily in areas where monkeypox is believed to exist, avoiding monkeypox requires adopting precautions to lower the chance of being exposed to the virus. Key preventive measures include:

- Avoiding Encounter with Wildlife: It is thought that animals, mainly rats and monkeys, are the source of the monkeypox.
- Washing hands frequently: Wash them frequently with water and detergent for a minimum of 20 seconds, especially after touching creatures, food from animals, or any potentially contaminated things [23].
- Respiratory Hygiene: Sneezing or coughing might cause droplets in the air that can spread the monkeypox. Important precautions, such as donning a mask in close proximity to an infected individual to lessen the likelihood of getting infected.
- Isolation and quarantine: Close contacts of suspected or verified patients ought to be confined to watch for signs. Whenever epidemics, healthcare authorities are essential in putting confinement and isolation procedures into place.
- Immunization: The general population cannot obtain a specific vaccine against monkeypox. The smallpox vaccination, however, has been demonstrated to offer some defense against monkeypox as well as may be given to healthcare professionals as well as others susceptible during epidemics [24].

1.1.6 Treatment of Monkey Pox:

The primary goal of treating monkeypox is to manage complications and provide supportive care to patients. Since monkeypox is a viral infection, there isn't a particular antiviral drug that may be taken regularly to treat it. Key components in treating monkeypox include:

- Patients with monkeypox should be isolated in order to stop the virus from spreading to others. Usually, the patient remains isolated until it's healthy[25].

- Hydration: It's important to be properly hydrated. Dehydration can occur in patients who have a fever, sweating, or skin lesions. Dehydration can be treated with intravenous fluids or oral rehydration treatments.

- ManagementofFever:Acetaminophen(paracetamol),anover-the-counterdrug,canbeused to lower fever and ease pain. However, because to probable side effects, non-steroidal anti-inflammatory medicines (NSAIDs) like ibuprofen should be avoided[26].

1.1.7 MalariaParasite

Malaria also remains one of the biggest global health problems, especially in sub-SaharanAfrica.ThisinfectiousdiseaseisspreadbyPlasmodiumparasitesandiscontractedbyhuman sthrough the bite of an infected female Anopheles mosquito. Malaria continues to be a seriousobstacle to socioeconomic development in affected areas despite significant efforts to controlit [27]. According to the World Health Organisation (WHO), malaria claims the lives of closeto 500,000 people annually, with children and pregnant women accounting for the majority ofthese fatalities. In addition to its initial negative effects on health, malaria has far-reachingeffects. The disease impedes economic growth in areas where it is endemic, which results inlower productivity, higher healthcare expenses, and a greater demand on already overburdenedhealthcare systems. Malaria also has an impact on education since ill kids are unable to attendthe lesson,whichfeedsthecycle ofplightandrestricted possibilities[28].

Although malaria cases have significantly decreased internationally, some areas still struggleto control the disease. The burden of malaria is made worse by elements including poorresources, a lack of effectiveinterventions, andrestrictedaccess to medical facilitiesinunderdeveloped and isolated places. Additionally, the dynamics of malaria transmission arecomplicatedbyshiftingenvironmentalcircumstancesandpopulationchanges[29].Epidemiological methods that consider variables like mosquito number, temperature, andhumidity are frequently used in conventional approaches for malaria forecasting. Althoughthese models have proved helpful, it's possible that they don't fully account for the complexrelationships and interactions that contribute to malaria transmission. Because of this, thesemodels' precision and dependability may differ, which reduces their capacity to forecast andstopepidemics [30].

In 2019, according to the World Health Organisation (WHO), there were approximately 228million patients with malaria worldwide in 2018, up from 231 million in 2017. Malaria isextremely common in the African continent, with an approximated 213 million instances ofthe disease identified in Africa, accounting for 93% of the overall instances, next to EasternAsia,whichcomprised5.4%ofoverallinstances,andtheEasternMediterranean,which accountedfor 2.1%ofallcases.MalariahasahighincidenceinSub-SaharanAfrica,withonlysix countries accounting for over half of the total number of malaria cases worldwide. ThesenationsareNigeria,whichmadeup25%, DemocraticRepublicofCongo(RDC),whichtalliedfor 12%, Uganda, which stood for 5%, and Mozambique, Niger, and Ivory Coast, which wereup4%each[31].

ML approaches provide the chance to revolutionise disease prediction and diagnosis becausethey are powered by sophisticated algorithms and computational skills. The goal of this studyistousemachinelearningtoanticipatetwoinfectiousdiseases:malariaandmonkeypox.Thesediseasesprovideparticulardifficultiesthatcallfor creativesolutionsbecauseoftheir intricacyandquicktransmission [32].

Table1:2:Symptomsofmalaria[33].

Symptoms	Percentage (%)
Fever	99
Chills	60-80
Fatigue	50-75
Muscle and joint pain	50-70
Nausea and vomiting	30-50
Cough	20-30
Diarrhea	10-30
Abdominal pain	10-20
Headache	70-90

Malaria,amajorthreat tohumanity’s health,istransmittedbyPlasmodiumfalciparumparasites, which are protozoan parasites. The parasites, which spread via the bites by infectedfemaleAnophelesmosquitos,haveacomplexlifecyclewithinpeopleandmosquitos.Afterthebites, theinfectiongoesintothebloodstream,theygothroughtheliver,wheretheyproliferatebefore being released back into the bloodstream, causing malaria symptoms such as fever andanaemia. Plasmodium vivax and P. ovale, for example, might hidein the liver and producerelapsesyearsaftertheinitialinfection.Asreportedby[33],feverisrecordedwiththehighest percentageinindividualsdiagnosedwithmalariaparasite,followed by headache.Othersymptomsarechills,fatigue,jointandmusclepainarerecordedwithhighsymptoms asshowninTable1:2.

1.1.8 Transmissionofmalaria:

Multiple phases are involved in the intricate life cycle of the spread of malaria. After the biteof aninfectedfemaleAnophelesmosquito on its host, sporozoites are injectedinto thebloodstream while the mosquito is feeding on the host's blood [34]. The liver is next infectedby thesesporozoites,where theybegintoproliferateandtransform intoschizonts.Thecirculatorystageoftheparasitesisstartedwhenthemerozoitesfromtheliverstagea rereleasedinto the blood. Merozoites enter red blood cells and reproduce asexually there. This causesinfected cells to burst, which starts the development of clinical symptoms. Some parasitesundergo sexual differentiation during this phase and develop into gametocytes, which can beconsumed byanothervectorafteranadditionalbloodmealasshownin Figure1:3 [35][37].

Figure1:3

Anophelesmosquito[37]

1.1.9 DiagnosisofMalaria:

Rapiddiagnostictests(RDTs),microscopicinspectionofbloodsmears,andmolecularapproaches such as polymerase chain reaction (PCR) are examples of common diagnostictechniques.Themosteffectivemethodtoidentify andclassify malariaparasitesisstillmicroscopic inspection, which involves a qualified technician looking at a blood smear undera microscope[36].

Figure1:4

Electricmicroscope[36]

By identifying specific parasite antigens in a blood sample, RDTs offer speedy results and areespeciallyhelpfulinenvironmentswithalackofresources.Inresearch,aswellasinsituationsof low parasitemia or suspected treatment resistance, molecular tools like PCR offer greatsensitivity and species discrimination. In places where malaria is endemic, especially, fast andsuitable treatment must be started as soon as possible. Accurate diagnosis is also crucial formalaria control and surveillance activities and could be used with the help of an electricmicroscopeas showninFigure1:4[36].

1.1.10 EpidemiologyofMalaria

A vector-borne transmissible illness, malaria displays a varied epidemiological pattern allthroughout the world. Tropical and subtropical areas are disproportionately affected, withAfrica bearing the greatest cost of the illness. Climate, geography, human behaviour, and thepresenceofsuitablemosquitovectorsaresomeoftheelementsthat affectmalariatransmission[38].

Figure1:5

Malariaparasiteepidemiology[38]

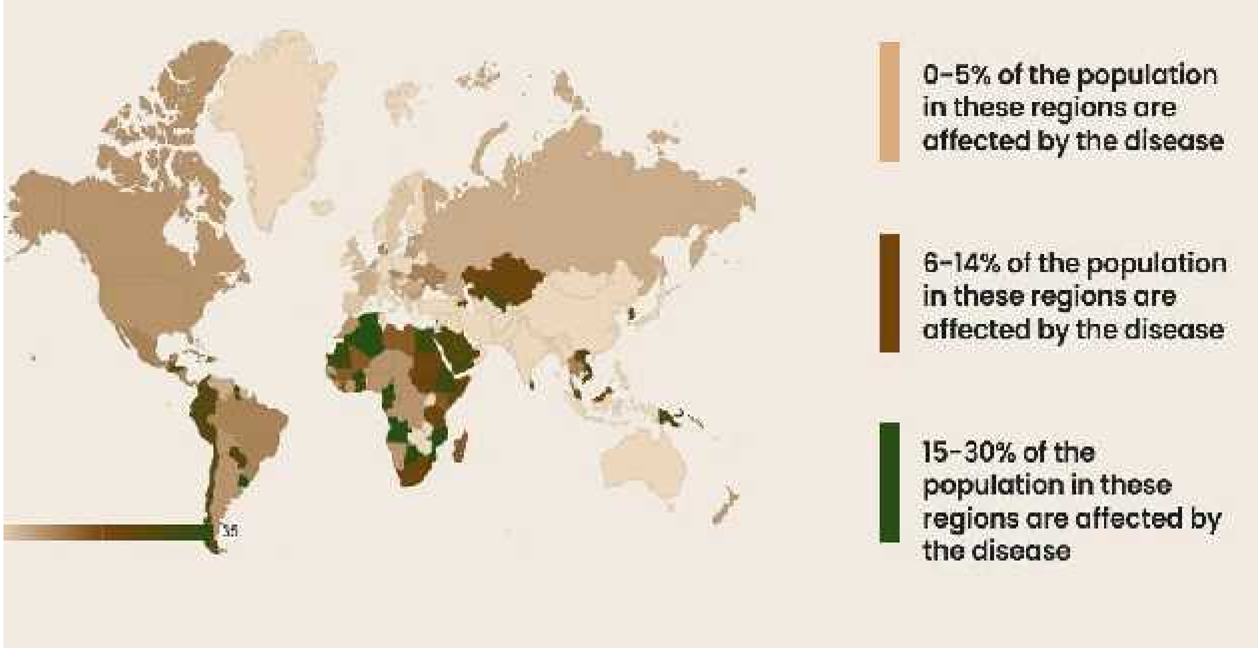


Figure1:5showstheglobewiththerecordofpopulationsmostlyaffectedbymalariainfection.The disease's prevalence typically peaks during the rainy season, when there are more placesformosquitoestospawn.Additionally,thereisahighnumberofdeathsfrommalariaeachyeara mongvulnerablegroups,includingchildrenunderfiveandpregnantwomen.Thedistributionof bed nets sprayed with insecticide, the use of antimalarial medications, and the creation ofpossible vaccines are just a few of the key advancements made in the management andpreventionofmalariainrecentyears [38].

1.1.11 PreventionofMalaria:

A multimodal approach is taken in the prevention of malaria with the goal of lowering bothdisease spreads and people's vulnerability to infection. The common adoption of insecticide-treated bed nets (ITNs) to prevent mosquito bites while people sleep, indoor residual spraying(IRS)togetridofmosquitoesinsidethehome,andchemoprophylaxiswithantimalarial medications for visitors to endemic areas are all part of this strategy. While programs foreducation emphasize the value of ITNs and quick treatment, ecological management andpreventioneffortsfocusonmosquitobreedingplaces.Womenwhoarepregnantmayoccasionally receive preventive care, and immunization is developing as a further preventiveintervention. The prevention of malariamustinclude early diagnosis and treatment, andcommunityinvolvementmakesthesemeasuresmoresuccessful[39].

1.1.12 Malariatreatment

In order to get rid of the Plasmodium parasites that cause malaria, antimalarial medicationsmust be administered. The Plasmodium species involved and patterns of treatment resistanceinfluencethetherapychoice.Chloroquinecouldbeutilisedincertaininstances,butcombinationmedicinesbasedonartemisinin(ACTs)are thepreferredtherapyformildfalciparum malaria. P. falciparum is frequently responsible for severe malaria, which requiresimmediate intervention with intravenous artesunate or alternative treatments [40]. Combiningantimalarial medications lowers the chance of drug resistance. The length of treatment varies,and it’s important to follow up. Pregnant women need to be treated differently, and additionalmedicationsmaybenecessarytostoprelapses.Inendemicareas,prompttherapyandpreventivetreatmentslikemosquitonetsareessentialto thecontroland eradicationofmalaria[40]. Conventionalmethodsofpreventing malaria,suchasusinginsecticidestokillmosquitoesandadministering antimalarialmedications,havehad some effectiveness.Due toissueslikepesticide resistance and the introduction of parasites with drug resistance, these approachesfrequently struggle with sustainability and execution. Improved artificial intelligence (AI)methodskeepofferingpreviouslyunheard-ofchancestobringabout arevolutionarychangeinthe direction of smart healthcare. The application of artificial intelligence (AI) approaches,including deep learning, machine learning, natural language processing, and other neuralnetworks, has greatly enhanced medicine commitment, compliance with therapy, and diseaseprediction,suchasmalaria.[41].

1.2 Statement oftheproblem

This thesis focuses on the prediction of malaria and monkeypox, using machine learningapproaches. In locations with limited resources, conventional testing methods associated learning (ML) techniques. This thesis aims to address these issues in order to improve ourcomprehension of ML applications in disease prediction and contribute to the creation oftrustworthysystemsforearlydiseaseidentification.

1.3 Aimoftheresearch

This research is motivated by the urgent need to address the ongoing problems caused byinfectiousillnesses,includingmalariaandmonkeypox,thatkeepravagingpopulationsalloverthe world, especially in settings with low resources. By utilizing machine learning (ML), theobjective is to improve prediction of malaria and monkeypox infections exceeding currenttechniques. Although ML has shown promise in a number of fields, its use in the realm ofinfectiousillnesspredictionisstillunchartedterritory.TheinvestigationandevaluationofMLalgorithms,suchasDecisionTrees,RandomForest,NaiveBayes,andLogisticRegression,aspotentinstrumentsforearlyandprecisediseaseidentificationisthemaingoalofthisstudy.

1.3.1 Objectives

The main goal of this study is to analyse and evaluate machine learning methods for malariaand monkeypox disease prediction. The study is led by the following particular objective inordertoaccomplishthis target.

- To assess the efficiency and performance of machine learning algorithms, such as LogisticRegression, DecisionTrees,RandomForest,NaiveBayes,andNaiveBayes,inforetellingtheriskofmalaria andmonkeypoxinfections.
- To investigate and put into practice data preparation approaches to improve the calibre andapplicability of input data for ML models. This covers feature engineering, feature selection,anddatacleansing.
- Tolookforwaystolessentheproblemsskeweddatasets,whicharewidespreadinhealthcaresettings , cause.Createmethodsfordealingwithclassimbalancetoguaranteeaccuratediseaseprediction.
- Tochooseandusetherightperformanceindicators,suchasPrecision,Recall,F1-Score,andSupport.

1.4 Significanceoftheresearch:

This study’s relevance has broad implications. By investigating machine learning methods forforecastinginfectionslikemalariaandmonkeypox,whichpredominantlyafflictvulnerable groups, it addresses a significant global health issue. This study offers the promise to protectlives, enhance healthcare outcomes, and lessen the demand for medical systems, particularlyin areas with few resources.By demonstrating machinelearning's useful applicationsinhealthcare, encouraging interdisciplinary cooperation, and spanning the gap between theoryand practice, it also promotes the science of machine learning. This study equips healthcareprofessionals, decision-makers, and academics to leverage the potential of machine learningfor the improvement of public health as well as underprivileged areas by offering practicalconclusionsandremediesforpracticalhealthcaresettings.

1.5 LimitationsoftheResearch:

- The availability of an extensive and well-annotated dataset for both training andassessingthestudycanimprovethe success ofthisstudy.
- Machine learning models' performance may be impacted by data imbalances, which isa frequentproblem.
- Classdisparitiesaredifficulttoaddress,andthestudy'sremediesmightnotbeadequatefortacklingextremeinequalities.
- The study's conclusions about the efficacy of machine learning algorithms might varydependingonthedataavailable.
- Themodelsgeneratedandevaluatedmaynotgeneralizeeffectively todifferenthealthcare contexts or locations with differing illness frequencies. It is essential tocomprehend the reasoning behind model selections, particularly in the healthcareindustry.

CHAPTER2

Literature Review

This chapter provides a thorough analysis of similar research that have previously beenundertakenandpublished,includingtheir findings,shortcomings,andoverallfindings.Thetheoreticalunderpinningsupportingthisresearch isalsoexplainedinthischapter.

2.1 OverviewofInfectiousDiseases(MalariaandMonkeypox)

The spreading of pathogenic organisms, such as microbes, viruses, parasites, or fungus fromone person to another, is the hallmark of infectious illnesses, a major public health concern.Both malaria and monkeypox are serious infectious diseases that have a global impact oncommunities.

The parasite Plasmodium causes malaria, an infectious disease spread by mosquitoes. Itpresentsasignificantpublichealthchallengebecauseitiswidespreadintropicalandsubtropical areas. Female Anopheles mosquitoes carrying the malaria virus bite humans tospreadthedisease.Redbloodcellsareinfectedbytheparasiteswhentheyenterthebloodstream,causingsymptomsincludingfever,chills,exhaustion,and,inseverecases,organdamage or death. Utilising bed nets sprayed with insecticide, taking antimalarial drugs, andcontrollingmosquitoes aresomepreventionoptions[42].

Monkeypox can affect both people and animals. It is mainly prevalent in nations in Centraland West Africa. The symptoms of monkeypox are similar to those of smallpox and include ahigh temperature, a rash, and cutaneous pustules. Transmission from person to person ispossible, mostly by exposure to bodily fluids or breathing droplets. Monkeypox is a problemeven though it is often less severe than smallpox, especially in areas where it is endemic.Preventive measures include maintaining adequate sanitation while avoiding coming intocontactwithwildlife[43].

The significance of managing infectious diseases, tracking, and treatment is shown by bothmalaria and monkeypox. In order to lessen the impact of these diseases and enhance globalhealth outcomes, public health initiatives, medical research, and community education arecrucial.

2.2 RelatedResearch

Advanced computational methods are revealed to be utilized in the prediction, diagnosis andpreventionofvariousdiseases.Thesecomputationalmethodsrangefrom statistical andmachine learning methods. Various disease and disease detection tools are evaluated andpredictedusingthesecomputationaltools[76].

According to their research, "Capacity Evaluation of Diagnostic Tests for COVID-19 UsingMulticriteria Decision-Making Techniques," Given the initial confusion around the choice ofdiagnostic tests during the early phases of the outbreak in Wuhan, China, Sayan, Murat, andcolleagues address the urgent necessity to compare and identify the most useful diagnosticapproach for COVID-19. Multi-criteria decision-making (MCDM) techniques, specificallyfuzzyPreferenceRankingOrganisationMethodforEnrichmentEvaluation(fuzzyPROMETHEE) and fuzzy Technique for Order of Preference by Similarity to Ideal Solution(fuzzyTOPSIS)wereusedintheirresearchtoassessdifferentdiagnostictestsforthecoronavirus disease [44]. The tests being performed include a chest CT scan, a polymerasechainreaction(PCR)analysis fordetectingviralnucleicacids,cellculture,aCoV-19antigenictest,aCoV-19antibodyIgMtest,aCoV-19antibodyIgGtest, andachestX-ray. TheresearchconcludesthatchestCTisamostefficientdiagnostictoolforCOVID-19usinglinguisticfuzzyscaleswithweightedfactorsbasedonprofessionaljudgments.Notably,theCOVID-19diagnosis was outperformed only by conventional diagnostic techniques for viral infections.Thisresearchstressesthesignificanceofadaptingdiagnosticapproachestoanation'sresourcesandoffersinsightfulinformationabout varioustestsfordiagnosiscombinations[44].

Different variants of cancer diseases are been predicted and detected using machine learningtechniques. Breast cancer screening is an important topic that Mustapha, Mubarak Taiwo, andcolleaguestackledintheirpapertitled"BreastCancerScreeningBasedonSupervisedLearningand Multi-Criteria Decision-Making," as the disease averages one woman dying from it everyminute. In order to increase survival rates, they emphasized the value of early diagnosis andsuggest a novel strategy that integrates computational intelligence with multi-criteria decision-making[45].Thebestmachinelearningalgorithmforearlybreastcancerdiagnosisisdetermined by comparing the performance of multiple supervised learning algorithms. Theydiscovered that the Support Vector Machine emerges as the best-performing model using thePreference Ranking Organisation Method for Enrichment Assessments, with a net outrankingflowof0.1022,confirmingitssuperiority.ThenaiveBayesclassifieristheleastrecommendedalternative, with K-nearest neighbour, logistic regression, and random forest classifier approach in assisting decision-makers in choosing the best machine-learning algorithm forbreastcancerscreening[45].

Inastudy,"EnsembleMachineLearningforMonkeypoxTransmissionTimeSeriesForecasting," Dada, Emmanuel Gbenga, and his colleagues tackled the urgent problem of theworldwide monkeypox outbreak, which has spread to more than 40 nations outside of Africaandhasbeendeemeda"publichealthemergencyofinternationalconcern"bytheWorldHealth Organization.Thestudiessuggestacomputer- andmachine-learning-assistedmethod toforecast the spread of the virus ratio considering the lack of information on the virus's results,risks, clinical manifestations, and distribution dynamics [46]. Adaptive Boosting Regression,GradientBoostingRegression,RandomForestRegression,OrdinaryLeastSquareRegression,Lasso Regression, and Ridge Regression are just a few of the machine learning methods usedin their study. They use performance indicators including Mean Square Error, Mean AbsoluteError,andRootMeanSquareErrortoassesstheeffectivenessoftheiralgorithms.Thefindingsshow that this Stacking Ensemble Learning (SEL) technique outperforms competing models,offeringinsightfulinformationaboutthethreatposedbymonkeypoxandassistingpolicymakersinimplementingthe appropriatemitigationmeasures[46].

Inasimilarwork,UzunOzsahin,Dilber,etalstudy,"Computer-AidedDetectionandClassificationofMonkeypoxandChickenpoxLesionsinHumanSubjectsUsingDeepLearningFramework,"addressestheurgentneedforpromptandprecisediagnosisofmonkeypox,apathogenicviralillness,inthemidstofacurrent epidemic.Theydrawattentiontothedifficultyindistinguishingbetweenchickenpoxandmonkeypoxduetotheresemblanceof the skin lesions, which might result in misdiagnosis and additional disease dissemination[47].Theyincorporateawell-trainedDeepLearning(DL)methodforearlylesionidentification and classification to address this issue. In their method, a two-dimensionalconvolutionalneuralnetwork(CNN)withthreeMaxPoolingstagesandfourlayersofconvolution is used. The suggested CNN algorithm outperforms competing DL approaches,outperforming them with a test accuracy of 99.60% and scoring 99.00% on the weightedaverage for precision, recall, and F1 score. With an accuracy of 98.00%, AlexNet leads allpreviously trainedmodels,whereasVGGNetperformedlesswith theaccuracy of 80%.Furthermore,thesuggestedCNNmodelisavaluabletoolfortherapidandpreciseclassificationofmonkeypoxadoptingcomputerizedpicturesof patient'sskinduetoitsgeneralizationandavoidanceofoverfitting[47].

In a study using the same dataset as in this study, Morang'a, Collins M., et al. classify es.

Their data set included haematological information on 2,207 people in Ghana who weredivided into groups according to whether they had a non-malarial infection (nMI), severemalaria(SM),oruncomplicatedmalaria(UM).Theyexperimentedwithsixvariousapproaches to machine learning before settling on one. They used a three-layer artificial neural network(ANN) for the multi-classification of UM, SM, and nMI. Additionally, binary classifiers werecreatedtofindthevariablesthatmaysetUMorSMapartfromnMI[48].

Their study's findings showed promise. The multiple classifications algorithm successfullydistinguishedbetweenmalariawithclinicalsymptomsandnMIwithtrainingandtestaccuracyof above 85%. The research identified the number of platelets, RBC (red blood cell) counts,lymphocyte counts, and percentages as the most effective classifiers for separating UM fromnMI,attainingatestaccuracyof0.801.Theclassifierdemonstratedagreattestaccuracyof 0.96 when separating SM from nMI, with both the mean volume of platelets and mean cellvolume being the main classifiers of SM. Platelet and RBC counts were crucial in identifyingUM,evenafterpotentialconfoundingvariableslikepatient ageandsamplelocationweretakeninto consideration, the study's Random Forest model utilised for verifying the results showed[48].

This study standsoutdistinctively amongtheotherstudiesmentionedaboveduetoitscombined focus on two serious infectious diseases malaria and monkeypox, thatprovideimportant public health issues. This research uses machine learning to forecast the prevalenceand spread rates for both malaria and monkeypox, whereas previous studies only focus on thedetection, categorization, or assessment of certain diseases. The urgent need for proactiveillness treatment and prevention is addressed by this distinctive viewpoint. Additionally, theincorporation of various machine learning models including Decision Trees, Random Forest,Naive Bayes, and Logistic Regression, as well as the assessment of performance metrics likePrecision, Recall, F1-Score, and Support, highlights the thoroughness of this work. This studyprovidesamorecomprehensiveandflexiblemethodforaddressinginfectiousdiseaseobstacles by investigating forecasting for two different infectious diseases and utilising a variety ofmachine learning techniques. This makes it uniquely exceptional in its range and conceivableeffectonpublichealth.

2.3 ArtificialIntelligenceinDiseasePrediction

Artificial intelligence (AI)is a subfield of computer science that aims to develop smartmachinesthatcancarryouttasksthattraditionallyrequirecognitiveability.Itincludesavariety ofdivisions,includingreinforcementlearning,deeplearning,computervision,naturallanguageprocessing,andmachinelearning.AIisevolvingquicklyandofferspotentialinmanyfields,including business,healthcare,self-drivingautomobiles,andotherfields[49].

In recent years, artificial intelligence (AI) has significantly improved medical and forecastingdiseases. Differentapplications of AI technology are being made to help with the earlyidentification, diagnosis, and treatment of illnesses. The following are some crucial elementsofartificialintelligenceindiseaseforecasting:

- **EarlyDiseaseDetection:**AIalgorithmsarecapableofanalyzinghugedatabases,suchaspatientrecords,testresults,andimagingdata,tospottrendsandanomaliessuggestive of ailment. This makes it possible to identify ailments such as cancer,diabetes,andheartdiseasepromptly[49].
- **MedicalImaging:**Deeplearning-basedtechniquesandotherAI-poweredimageprocessingtoolshaveshownconsiderablepromiseinidentifyingdiseasesusingX-rays,MRIs, andCTscans.Thesetoolscanhelpradiologistsidentifyanomaliesandincreasethe precisionoftheirdiagnoses.
- **Genomic Analysis:** A person’s vulnerability to genetic illnesses and associated riskfactorscanbepredictedusinggenomicdataanalysisperformedbyAI. Additionally,itcan aid in the discovery of prospective targets for the development of new drugs andcustomizedtherapies.
- **PredictiveAnalysis:**Usingpredictivemodelling,AIcandetermineaperson’slikelihood of contracting a particular disease based on their genetics, lifestyle choices,andmedicalhistory.Offeringindividualizedpreventionandinterventionmeasures withthe aidofthisinformationcanbenefitpatients[50].
- **Wearabletechnologyandhealthsensors:**WearabletechandhealthsensorsthathaveAI built in can continually monitor a person’s health indicators including heart rate,blood pressure, and glucose levels. To assist in managing chronic conditions andidentifyingnewhealthissues,thesegadgetscandeliver dataandalarms inreal-time.
- **Analysis of Electronic Health Records (EHR):** EHR can be analyzed by AI to findindications of trends in patient data. This can help in epidemic prediction, healthcareprocessoptimization,andraising thestandardofcareingeneral.
- **DrugDevelopment:**Byanalyzinglargedatasets,modelingchemicalinteractions, andforecasting possible medication candidates, AI-driven technologies are hastening thedrugdiscoveryprocess.Newtreatmentsfor avarietyofdiseasesmayresultfromthis.

- **Remote monitoring and telemedicine:** Chatbots and artificially intelligent assistantswith AI capabilities are used to give medical recommendations, respond to inquiries,and keep an eye on patients from a distance. These innovations can improve patientaccesstocareandlightentheloadonmedicalinstitutions[51].

Figure2:1

HierarchicalrepresentationofA.I,ML andDL[52]

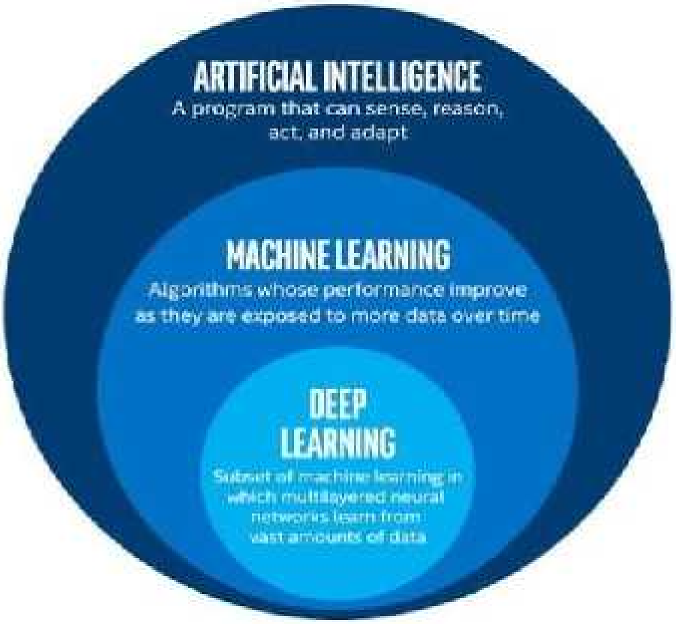


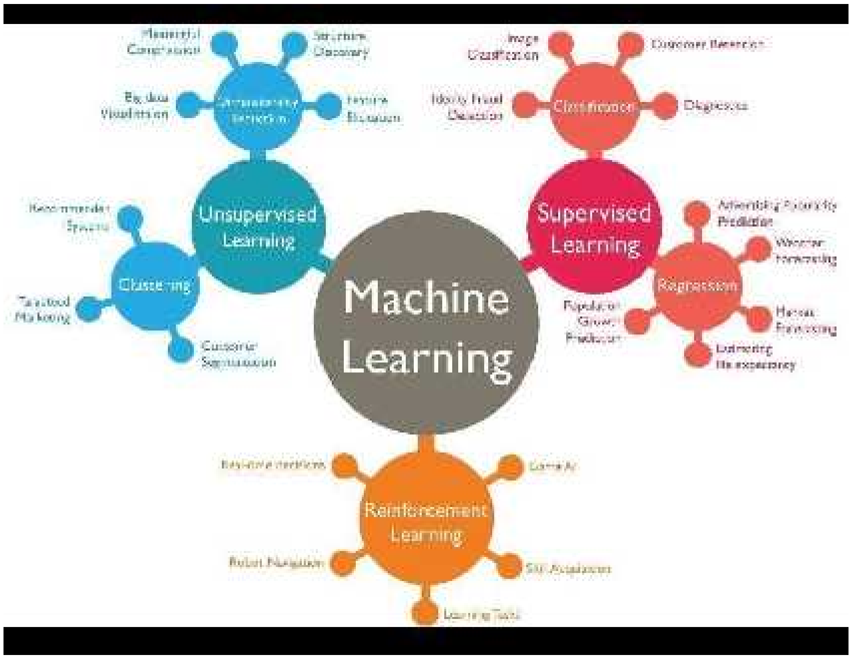
Figure 2:1 shows the hierarchical illustration of AI, ML and DL. This figure defines anddifferentiates theseprocessesandtheirimportancetothe fieldofscience.

2.3.1 MachinelearninginDiseaseprediction

Machinelearningis asubfieldofartificialintelligencethatfocusesoncreatingtechniquesthatallowmachinestogain knowledgefrominformationandmakeforecastsorjudgments.Applicationsincludenatural language processing,computervision,and recommendationsystems. It includes models being trained on facts to identify trends and connections. Byutilizing data-driven insights, machine learning algorithms, such as supervised, unsupervised,and reinforcement learning, have revolutionised the process of problem-solving and makingdecisions,makingitanessentialpartofcontemporarytechnologyandAI[53].

Figure2:2

Machinelearning sub-set[54]



In order to uncover trends and components linked to diverse medical illnesses, machinelearninganalysesmassivedatasets,thatarecrucialfor forecastingdiseases.Ithelpswithtimelydiagnosisandisutilizedtoanalyseriskvariablesfordiseasesandforecastthechanceofillnessincidence.Todeliverindividualizedriskevaluationsandsuggestprecautionarysteps,predictive algorithms can make use of patient details, genomic data, and medical histories.Furthermore, they are used in medical imaging to diagnose diseases and may analyse patientinformation to forecast disease outbreaks and optimise the use of medical resources, all ofwhich help to provide more successful disease prevention and control techniques as shown inFigure2:2[54].

2.3.2 Supervisedmachinelearning

In the field of supervised machine learning, techniques are trained using labelled datasets thatcombine input data with appropriate output labels. The technique ought to establish how tonavigate from input to the result so that it can make predictions or categorize previouslyunexploredmaterial[55].Itisfrequentlyusedfortaskslikeclassification,whereitlabelsor categorizes data, and regression, where it forecasts continuously changing numerical values.Theeffectivenessofsupervised learning,whichisused inavarietyoffieldssuchasprocessingnatural language, computer vision, and forecasting, depends on having high-quality labelleddata. The trained model's effectiveness is largely determined by its capacity to generalize tonewdata,therefore properdatapre-processing,choice ofmodels,and assessment[55].

Many different machine learning algorithms fall under the category of supervised learning,mostlydependingonthetypeofresultvariable.Transferringdatatopredetermined categoriesistheprocessofclassification;abinarysystemhastwocategories,whilemulti-classclassification has many [56]. Ordinary regression deals with ordered, distinct outcomes whileregressionforecastsnumericalvalues.Datamightconcurrentlybelongtonumerouscategories thanks to multi-label classification. Datasets having skewed class distributions are addressedvia improper classification. Rare, anomalous occurrences are found by anomaly detection.These supervised learning approaches provide multiple options for various real-world issues,eachrequiringcertainmethodsandstrategiesdependingontheattributesoftheoutputvariable [56].

- **Classification:**Theobjectiveofclassification,asubsetofsupervisedmachinelearning,istocategorizeorclassifyinputdatapoints.Buildinganalgorithmonalabeleddatasetthatcontainsinputattributesandassociatedclass labelsisrequired.Thealgorithmthatwas trained then uses the trends it discovered during training to make forecasts on thekind or group of brand-new, unobserved data points. The goal of classification tasks isto divide data into separate groups or classes depending on its properties or features.Examplesofpopularclassificationtasksincludemaliciousemailidentification,recognitionofimages,sentimentevaluation,andmedicalevaluation.Commonmetricsforassessingtheeffectivenessofalgorithmsforclassificationincludeprecision,recall,precision,and F1-score[57].
- **Regression:**Adiscretenumberorquantityistobepredictedusingtheinputcharacteristics in regression, a form of supervised machine learning. On a labeleddataset, where every point of data contains input variables and associated goal values,a model is trained. In order to estimate the desired value of fresh, unforeseen datapoints,thetrainedregressionmodellearnstorecognizepatternsandcorrelationswithinthedata.Regressiontasksarecommoninmanydifferentindustries,suchasevaluatinga patient's blood pressure based on health data or predicting house prices based onfactorslikesquarefootageandlocation.Meansquarederror(MSE),meanabsolute

error(MAE),andR-squared(R2)arecommonmetricsforassessmentformodels thatuse regression[58].

2.3.3 UnsupervisedMachinelearning

Machine learning that uses unlabelled data and focuses on finding trends, frameworks, orcorrelations amongst the data with no predetermined output labels is known as "unsupervisedlearning." It entails activities like clustering, in which data points are categorized according tocomparable characteristics, reduced dimensionality methods for streamlining complicateddata, detecting anomalies for locating outliers, and estimation of density for calculating thedistribution of data. In numerous fields, such as disease segmentation, picture evaluation, andnaturallanguageprocessing,unsupervisedlearning isusefulforuncoveringhiddeninsightsinhuge,unstructureddatasets[59].

- **Clustering:**Unsupervisedmachinelearningmethodssuchasclusteringcategorizedatapoints into groups or subgroups that have shared characteristics or trends. Lackingprevious understanding of the groups or labels, the main objective of clustering is tofind innate patterns or clusters inside the data. The use of K-means hierarchical areexamplesof typical clusteringtechniques.Examples of clusteringinclude textualclustering in natural language processing, image segmentation in computer vision, andclient segmentation in marketing. By exposing deeper trends and interactions betweendata points insidevarious clusters, aidsin theorganization and understanding ofcomplicateddatasets[60].

2.3.4 DeepLearning:

Deep learning is A branch of neural networks, or a component of artificial intelligence (AI)that involves modelling and resolving complicated issues using artificial neural networks,particularlydeepneuralnetworks.Thesedeepneuralnetworksareusedtointerpretandmodifydata using numerous layers of interrelated nodes (artificial neurons). Deep learning's keyattributes include:

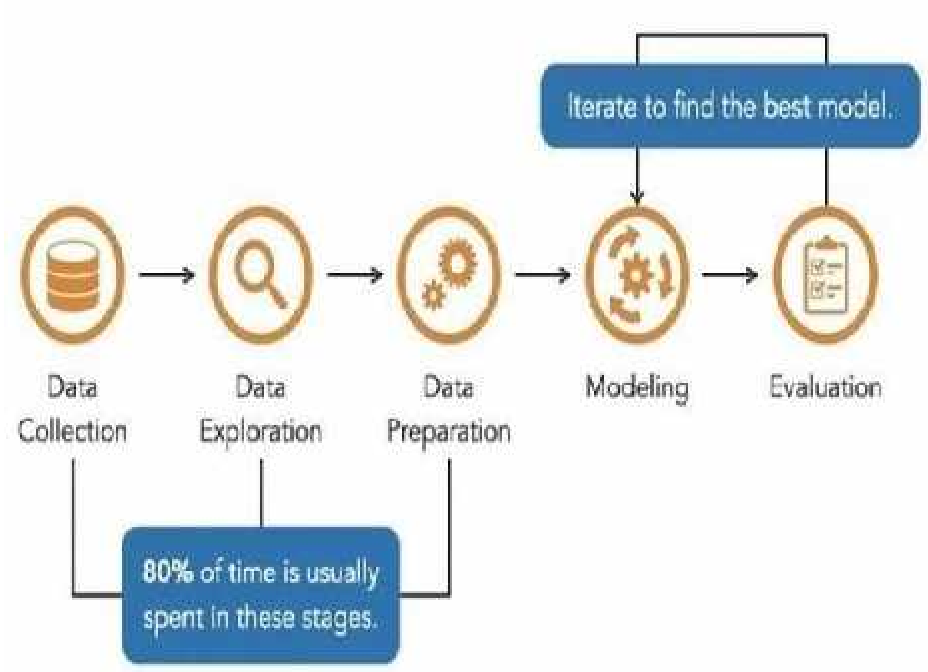
- **Numerous layers:** The input layer, several layers that are hidden, and the output layerare just a few of the layers that make up deep neural networks. The network may learnprogressively more abstract and complicated aspects from the incoming data thanks totheselayers[61].

- **Hierarchical Features:** Deep learning models naturally pick up on hierarchical datarepresentations.Higherlayerscapturemoreabstractandcomplicatedpatterns,whereaslowerlevelsfocus onmorefundamentalelements.
- **Deep learning models** learn from enormous volumes of labeled data through data-drivenlearning.Throughaprocedureknownasbackpropagation,errorsaretransmittedretrogradeviathenetworkinordertoamendbiasesandweights,allowingthemtogrowandchangetheirowninbuiltproperties [62].
- **Numerous Uses:** Deep learning has been used in a wide range of fields, includingcomputer vision (for example, image recognition), natural language processing (forexample,languagetranslationandsentimentanalysis),speechrecognition,recommendationsystems,and autonomoussystemslike self-drivingcars.
- **Deep Learning Frameworks:** To render it simpler for scientists and programmers tocreateandtraindeepneuralnetworks,anumberofdeeplearningframeworks,includingTensorFlow,PyTorch,andKeras,havebeencreated[63].

2.4 TheTheoreticalFrameworkoftheDissertation

Figure2:3

Procedural Flow



AsshowninFigure2:3anorganizedmethodologyisused inthetheoreticalframeworkfortheresearch on diseasepredictionemployingmachinelearning,andappropriatedatasets onailmentslikemalariaandmonkeypoxarecollectedaspartofthisprocess.Datapre-processingandcleaningguaranteetheaccuracyandreliabilityofthedataasdoneby[64].Forthepurposeof developing and accessing models, data is then divided into sets for testing and training. Tocreate forecasting models for the ailments, a number of machine learning algorithms are used,such as Decision Trees, Random Forests, Logistic Regression, and Naive Bayes. The best-performing model is chosen for implementation in actual medical facilities after the model'sefficiencyis evaluated,similarto[65].

2.4 MachineLearningModels

Thediseasepredictingstudymakesuseofarangeofmachine-learningframeworks,includingDecision Trees, Random Forests, Logistic Regression, and Naive Bayes. These models werechosen based on a variety of qualities and propensities for predicting illnesses like malaria andmonkeypox. Naive Bayes is effective for particular data types, while Logistic Regression issimpleandDecisionTreesgiveinterpretability.RandomForestsimprovepredictionaccuracy.Withthespecificdiseasepredictiontasksathandaswellasthedataset'sdistinctivecharacteristics in mind, the study examines and assesses the results of different models todiscoverwhichoneisthemostefficientforforecastingdiseases, similartostudyby[66].

2.4.1 DecisionTree

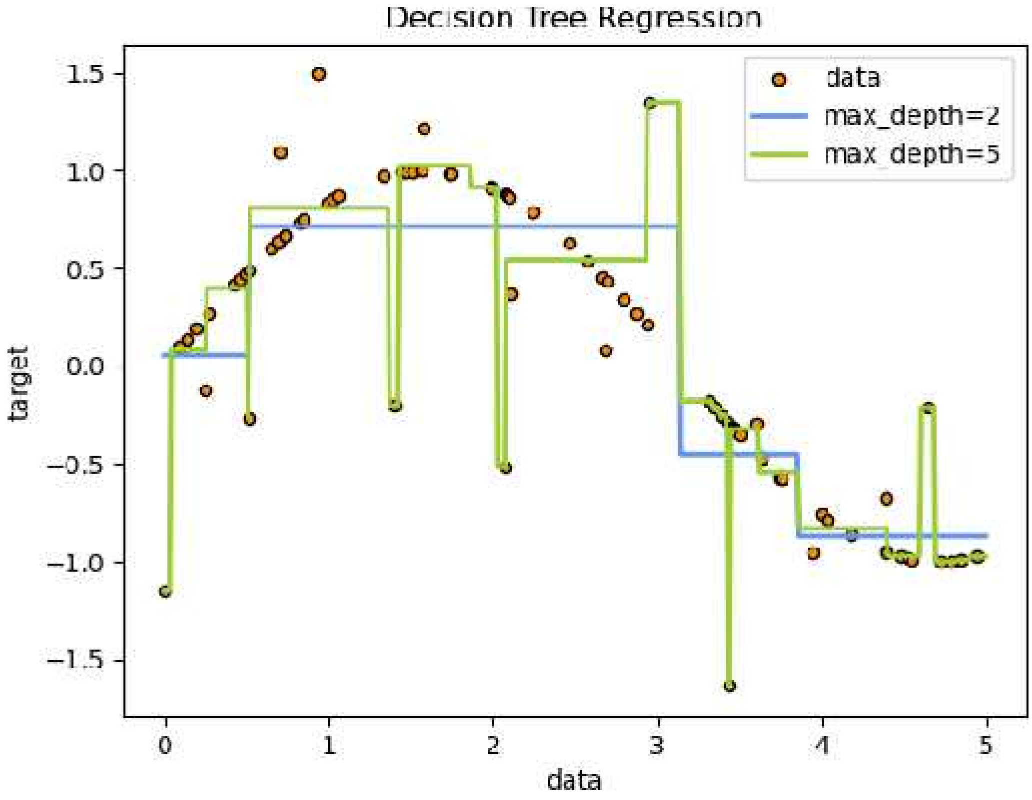
Decision Trees (DTs) are an irregular supervised learning technique for both regression andclassification. By learning straightforward decision rules derived from the data attributes, theobjective is to build an algorithm that forecasts the value for a target variable. A tree can bethoughtofas afragmentedconstantestimate[67].

A decision tree is a machine algorithm and visualization utilized in tasks such as regressionand classification. It comprises networks indicating judgments made in response to particularfeatures, branches that produce offspring nodes, and leaves offering ultimate forecasts orclassifications. Decision trees are accessible and ideal for activities requiring decision claritybecause they use metrics like entropy as well as data gain to calculate feature splits. Althoughthey can be inclined towards overfitting, pruning strategies assist the tree becomes morestraightforward.DecisionTreesareusefulinmanydifferentdomainsandarefrequentlyused as the foundation of ensemble techniques like Random Forests. They are useful tools forcomprehendingandpredictingdata duetotheirclarityandinterpretation[67].

Decision trees over-fits if the optimum depth of the tree (managed by the max_depth option)issettoohighbecausetheyacquiretoomanysmallfeaturesfromthetrainingdataasshowninFigure2:4.

Figure2:4

DecisionTreeRegression(referencescikit-learn) [68]



2.4.2 RandomForest:

A random forest, which functions as a meta estimator, uses aggregating method to improveforecastingaccuracyanddecreaseoverfittingbyfittingseveraltypesofdecisiontreeclassifiersson various subsamples of the dataset. The max_samples option controls the size of the sub-sample when bootstrap=True (the default); in contrast, every tree is built using the wholedataset when bootstrap=False [69]. A large number of decision trees are built during thetraining phase of the random forest approach, which is used for classification, regression, andother tasks. The category that the majority of the trees chose is the result of the random forestforclassificationproblems.Theaverage,ormean,forecastforeverypersoninthetreeisgiven back for the regression tasks. The tendency of decision trees to overfit their training set iscorrectedbyrandomdecisionforests.Althoughrandomforeststypicallysurpassdecisiontrees,the yarelessaccurateincomparisontotrees.However,theireffectivenessmaybeimpactedbydata peculiarities [69].

2.4.3 NaiveBayes

A family of basic "probabilistic classifiers" known as naive Bayes classifiers in data analysisis built on the application of Bayes' theorem with solid (naive) independence assertions withattributes (see Bayes classifier). They are some of the most straightforward Bayesian systemmodels, but when used in conjunction with kernel density estimation, these can attain greatlevels of efficiency [70]. The total quantity of parameters required for naive Bayes classifiersis linear in a given amount of data points (features/predictors) in the learning issue, makingthemhighlyscalable.Insteadofusingthecostlyiterativeestimation,asisthecasefornumerous otherkindsofclassifiers,maximum-likelihoodlearningcanbeperformedbysimplyassessingaclosed-formopinion,thatrequiresalineartimeframe[70].

2.4.4 LogisticRegression

Thelogisticapproach,oftenknownasthelogitmodel,isananalyticalframeworkthatestimatesthe likelihood of an occurrence occurring by making the outcome's log odds a linear sum ofseveral independent variables. In regression analysis, the technique of logistic regression isusedtoestimatealogisticmodel'sfeatures.Inmathematicalterms,inlogisticregressionmodel,there is only one binary dependent variable that's coded by a signal parameter, where bothvalues are labeled "0" and "1", whilst the variables that are independent are able to be either abinaryvariableoravariable thatis continuous[71].

2.4.5 EvaluationMetrics

Techniquesformeasuringtheefficacyofpredictivealgorithmsfortasksrequiringtheclassification of data into specific groups or subcategories are called classification metrics forassessment [72]. The F1-Score, which reconciles both recall and precision, the specificity aswell as false positive percentage, useful for distinguishing adverse instances of the ROC andprecision-recall arcs, which visualize the accuracy of models across differentlevels, andCohen'sKappa,usedtoevaluateagreement,areexamplesof typicalmetrics.Accuracydetermines general correctness; precision evaluates the number of accurate positive forecastsrecall examines the capacity to recognise all true positives; recall determines the capacity positives and false negatives all influence the selection of metrics, which frequently calls foranamalgamationofmetricsforathoroughevaluation[72].

2.4.5.1 ConfusionMatrix

A confusion matrix is a data table that summarises how well a machine learning modelperformed in categorization assignments. It offers a full description of the model's forecastsand actual results, allowing for a thorough assessment of the model's effectiveness [73]. Theconfusionmatrixnormallyincludesthefollowingfourcrucialparts:

True Positives (TP) are situations in which the algorithm accurately forecasts the class ofpositives,foraswhenitcorrectlyidentifiespatientswithacondition.

TrueNegatives(TN)aresituationswherethealgorithmaccuratelyforecaststhenegativeclass,foras whenitproperlyidentifieshealthypeople asbeingfree ofthe disease.

FalsePositives(FP)aresituationswhereanalgorithmmistakenlyforecastedaclassaspositivewhile the actual class was a negative one (for example, identifying a healthy person as havinga medicalcondition.

FalseNegatives(FN)aresituationswhenthealgorithmforecastedthewrongclass—anegativeone— while the real class was a positive one. An example of a False Negative (FN) occurswhen the algorithm fails to identify an illness in the patient who actually possesses it. Thismistakeis alsoreferredtoas aType IIerror[73].

2.4.5.2 ClassificationAccuracy

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$$

Thecapacityofthealgorithmtoaccuratelyclassifycasesintospecifiedcategoriesisquantifiedby the basic assessment statistic known as classification accuracy in the field of machinelearning.Itcalculatesthepercentageofthetotaldatasetwhoseinstancesthealgorithmcorrectlyforecasts. This measure is simple to understand and is frequently represented as a percentage.Assuming groups in the dataset have equal values and incorrect classification rates are aboutequivalent for all classes, accuracy is useful for evaluating model performance. However,accuracyhasseveralrestrictions.Concentratingonlyonaccuracycanbedeceptiveinsituationswhere there are class imbalances or where the effects of false positives and false negativesdiffer dramatically.For a more complete analysis, itis imperative tocombine accuracyalongsideadditionalmetricssuchas precision,recall,theF1-Score,orROCcurves[74].

2.4.6 ClassificationReport

- **Precision:**

Precision measures an algorithm's capacityto prevent false positives bycomparingthepercentageoftruepositiveforecaststoallpositiveforecasts[75].

$$Precision=\frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

- **Recall:**

Recallmeasureshowwellanalgorithmcanidentifyallpositivesbydividingtheratiooftruepositiveforecasts bythetotalnumberofpositive cases:

$$Recall=\frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

- **F1-score:**

The F1-Score is an extensively utilized classification metric thatincorporates recall andaccuracy to offer an objective evaluation of a machine learning algorithm's performance,particularly when working with unbalanced datasets or where the price of false positives andfalsenegativesvary.Itis especially helpful whenbalancingthenecessity of accuratelyrecognizingallactualpositives(recall)withthe need tomakeprecise positiveforecasts.

$$F1-score=\frac{\mathrm{2x(precisionxRecall)}}{\mathrm{Precision+Recall}} \quad [75]$$

CHAPTER3

Methodology

The technique used in the study is carefully described in this chapter, which also provides astep-by-step overview of the process by which the research was carried out. It provides anextensive description of the data analysis procedure and painstakingly outlines the steps takenfor data acquisition. This chapter essentially lays out the organized architecture that served asthe research’s direction throughout the stages of data collection, analysis, and interpretation. Itacts as a road map for comprehending how the study was carried out and how the study’sconclusionswerearrivedat.

3.1 Datacollection andStudyPopulation:

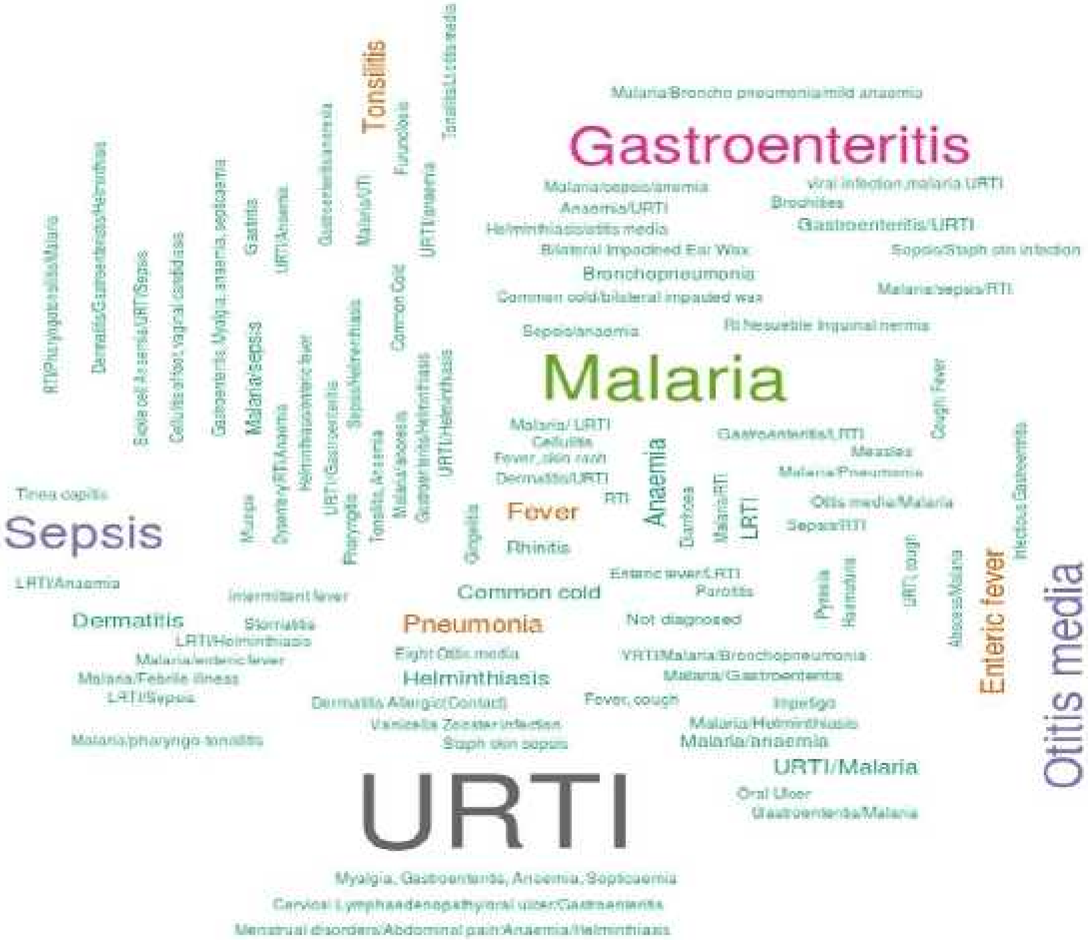
The malaria dataset implemented in this study was obtained from an online resource with freeaccess that is accessible through Kaggle.com. The dataset consists of 2,207 cases of malariaparasitesymptoms.Originally,CollinsM.,Morang’a[48],andotherresearchersassembledthisdatacollectionasacomponentofaresearchstudy.Theirresearchcenteredonutilizingmachinelearningmethodstocategorizeclinicalmalariaoutcomesusinghematologicaldataasfeatures.It’s important to note that the initial dataset, on which this study is based, was gathered at theNavrongoHealthResearchCentre oftheRepublicofGhana.

Similarly, the monkeypox dataset was also obtained from Kaggle.com. There are 11 variouscharacteristics or symptomic factors in this dataset, which consists of 25,000 instances. Thefeatures such as PatientsID, systemic illness, rectal pain, sore throat, penile Oedema, orallesions, solitary lesion, swollen tonsils, HIV infection, sexually transmitted infection and themonkeyPoxasthetarget.

The majority of the 150 cases had upper respiratory tract infections (URTIs), followed bymalaria, gastroenteritis, sepsis, otitis media, and fever. The rest of the determinations weremadeonfewerthantwooccasionseachperyearasseeninFigure3:1.

Figure3:1

Cloudrepresentingclinicalsymptoms

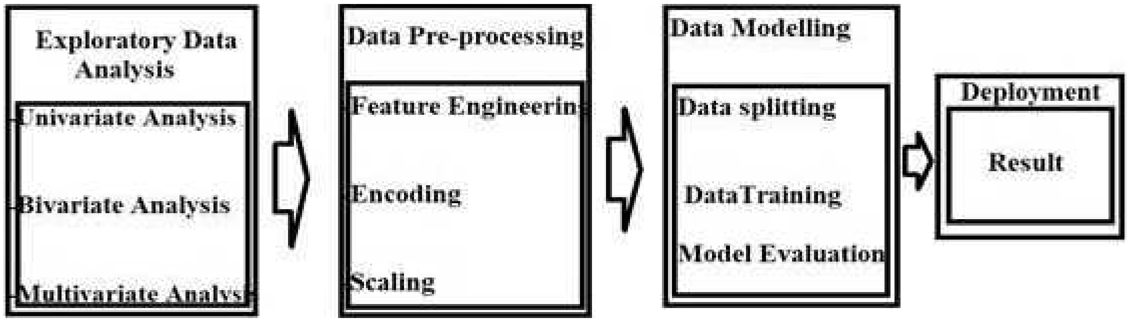


3.2 MachineLearningPipeline:

In a typical machine learning pipeline, there are several clear steps. Data collection is the firststep, then comes data preprocessing, which involves cleaning, manipulating, and separatingthe data. Toproduce accurateinputsfor the algorithm,featureengineeringisessential.Following model selection, the proper method is selected, and the resulting model is thentrained.Theframework’ssettingsareoptimizedbyhyperparametertuning,andperformanceis measuredthroughevaluationmetrics.Testingonadifferentdatasetconfirmsgeneralization.Ifthe model performs well, accessibility approaches may be used, and it is then deployed. Itseffectivenessiscontinuouslymonitoredandmaintained,andfindingsarereportedanddocumented. Using a feedback loop, a model can get better over time. This process offers adisciplined method for creating and implementing machine learning techniques for a variety oftasks as shownonFigure3:2[77].

Figure3:2

Machinelearningpipeline

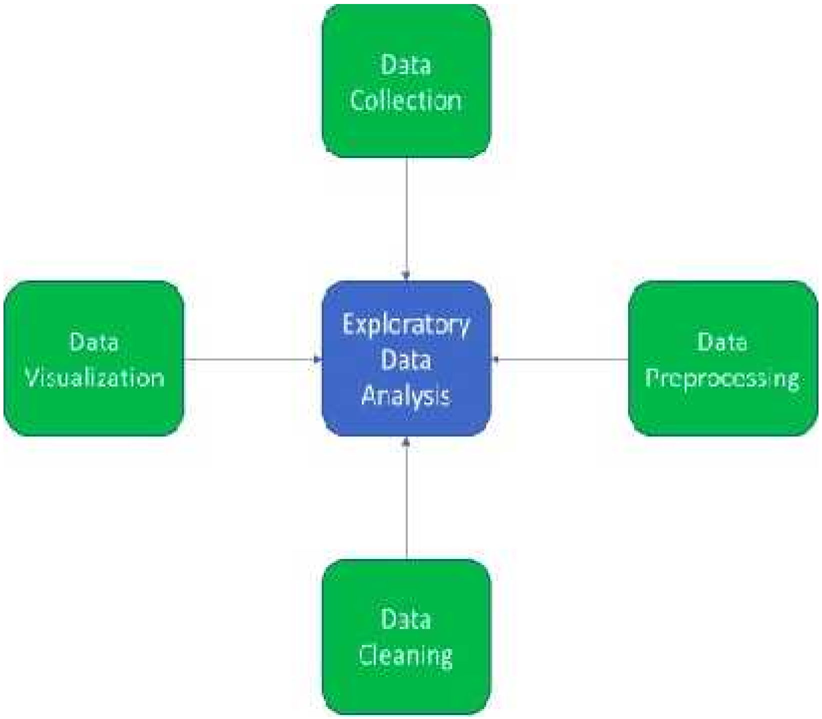


3.3 ExploratoryDataAnalysis

In order to get ideas and comprehend the data’s fundamental properties, exploratory dataanalysis (EDA), a crucial early-stage technique in data analysis, entails gathering, cleaning,visualizing, and summarising the data. Data cleaning to assure data quality, data visualizationtoproducevisuals,andthecomputationofdescriptiveanalysisareallincluded.EDA investigatestheconnectionsbetweenparameters,looksforanomalies,andincludeshypothesistesting. EDA frequently involves repetitive processes, and feature engineering and validation arecrucial elements. Its conclusions offer a sound basis for decision-making and subsequent datamodelling indata analysisandmachinelearninginitiativesasshowninFigure3:3 [78].

Figure3:3

ExploratoryDataAnalysis[78]



- **UnivariateAnalysis:**

Univariate analysis is an Exploratory Data Analysis technique designed to examine just onevariable in a dataset. Its main objectives are to comprehend the distribution of the parameter,detect outliers, and summarise statistics and visualizations to represent the features of thevariable. In this study, each variable in the two datasets was reported using the univariateanalysis[79].

- **BivariateAnalysis:**

Bivariateanalysisisanexploratorydataanalysismethoddesignedtoinvestigateandcomprehendtheinteractionbetweentwovariablesinacollectionofdata.Itmakesuseoftools includingcrosstabulation, regressionanalyses,analysisofcorrelation,scatterplots,andtestinghypotheses.An understandingofrelationships,trends,anddependenciesbetween twoparameters can be gained by bivariate analysis, which also offers important information onhowvariationsinonevariableaffectchangesinanother.Identifyingcorrelationsanddirectingmorestatisticalormodellingresearch,itisavitalstageinthedataanalysis process.

- **Multi-variateAnalysis:**

Multivariate analysis is an exploratory data analysis technique used to analyse and figure outthecorrelationsbetweenseveralvariablesinadataset.Multivariateanalysisconsidersatleast threevariablesasopposedtojustoneortwothatunivariateandbivariateanalysesdo,allowingfortheexaminationofintricateconnectionsandrelationships.Multivariateregression,principal component analysis (PCA), factor analysis, cluster analysis, canonical correlationanalysis (CCA), the modelling of structural equations (SEM), as well as additional methods,are frequently used in multivariate analysis. This method can be used to findcomplicatedpatterns,predictthefuture,andinterpretcomplicatedstructuresofdatainavarietyof disciplinesaswellas studyareas.

3.4 DataPre-processing:

Data pre-processing is essential to machine learning studies that improves the performance ofdata.Inthisstudy,thedatasetunderwentpre-processingusingPython(version3.8)inaJupyternotebook.Inparticular,neitherthedatasetsfor malarianormonkeypoxcontainedanydatathatwasmissingorzerovalues,eliminatingtherequirementformethodstoaddressthem,includingelimination or imputation. Additionally, the datasets were adequately balanced, thus neitheroversampling nor under sampling techniques were required. To guarantee there was enoughtrustworthy data for the inquiry, anomalies in the data set have been substituted with suitablevalues. By carefully handling missing values and outliers, it is possible to lessen redundancy,ambiguity, and noise, which eventually aids in the creation of reliable and effective predictivemodels.

- **DataCleaning:**

Data cleaning, commonly referred to as data cleansing, is a crucial stage in the pre-processingof data.Adataset'saccuracy,completeness,andsuitabilityforanalysisareitsprimaryobjectives. Managing missing values, duplicate data, and outliers, standardizing data formats,resolving discrepancies, resolving encoding problems, confirming data kinds, and imposinglimitationsaresomeofthemajorjobsindatacleaning.Maintainingdataqualityandreliabilityis essential for allowing appropriate and insightful data analysis and modelling, which in turndepends oneffectivedatacleansing[78].

- **Encoding:**

Encodingistheactoftransformingtextualorcategorydataintoanumericalformatappropriatefor machine learning and data analysis. It is crucial since numerous algorithms for machinelearning need input in the form of numbers. Label encoding, single-hot encoding, binaryencoding, frequency encoding, target encoding, ordinal encoding, and embedding are only afewexamplesofthemanyencodingtechniquesavailable.Thetypeofencodingtechnique chosen relies on the data’s structure and the particular machine learning algorithm being applied. Effective model learning and predictions are made possible by proper encoding, which guarantees an accurate representation of categorical information[80].

- **Scaling:**

Scaling is a fundamental part of data preparation in machine learning and involves transforming numerical data into a stable range or scale. The two most used scaling methods are standardization and Min-Max scaling (normalization). Using min-max scaling, attributes are condensed to a narrow range, usually between 0 and 1, while still keeping linkages between data points. After standardization, which converts variables to have a mean of 0 and a standard deviation of 1, data is centered about zero and given a unit variance. The precise requirements of the learning method being used dictate the scaling methodology that is used. Scaling ensures that each feature contributes equally to the modelling process, improves model performance, and stops larger-scale features from controlling the study[81].

CHAPTER4

Results

The outcomes of the methods used to select and evaluate two datasets relevant to malaria andmonkeypoxarecoveredinthis chapter.

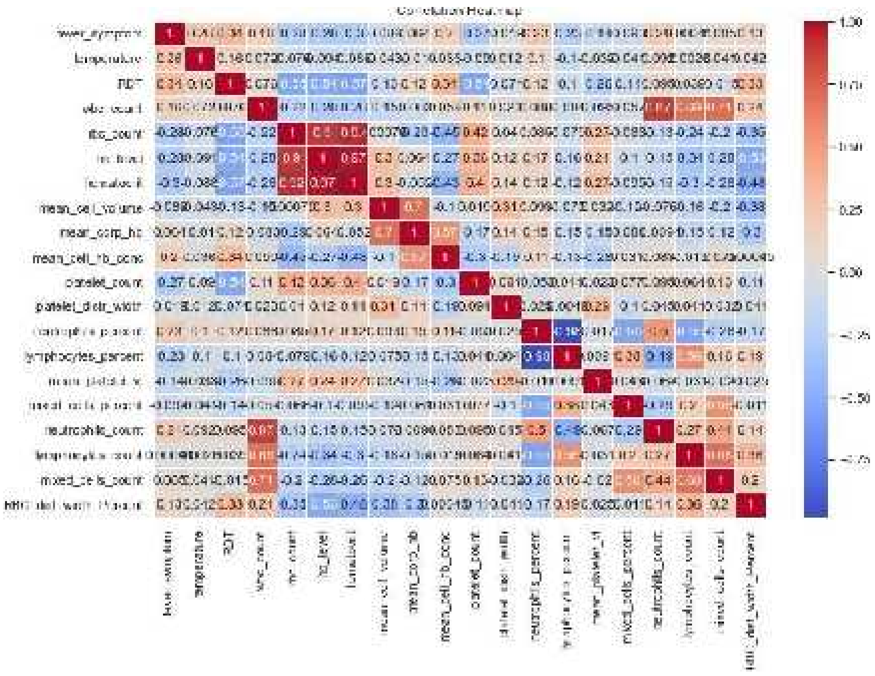
4.1 PredictiveModelsforMalaria:

4.1.1 Variables

This study variables include 20 variables, with both numerical and categorical data types. Thevariablesaretemperature,RapidDiagnosticTest(RDT),whitebloodcellcount,redbloodcellcount,hemoglobinlevels,hematocrit,mean_cell_volume,meancorpuscularhemoglobin,meanconcentrationofhaemoglobininbloodcells,plateletcount,plateletdistributionwidth,percentage of neutrophils, percentage of lymphocytes, mean platelet volume, percentage ofmixedcells,count ofneutrophils,count oflymphocytes,countofmixedcells,anddistributionwidth of red blood cells. Furthermore, Microscopy is the categorical variable, which is thetarget.AsshowninFigure4:1,thecorrelationheatmapillustratestherelationshipbetweentheindependentanddependentvariables.

Figure4:1

CorrelationHeatmapofmalariadataset



4.1.2 ExploratoryDataAnalysis

4.1.2.1 UnivariateAnalysis

Figure4:2

UnivariateAnalysisofMicroscopy

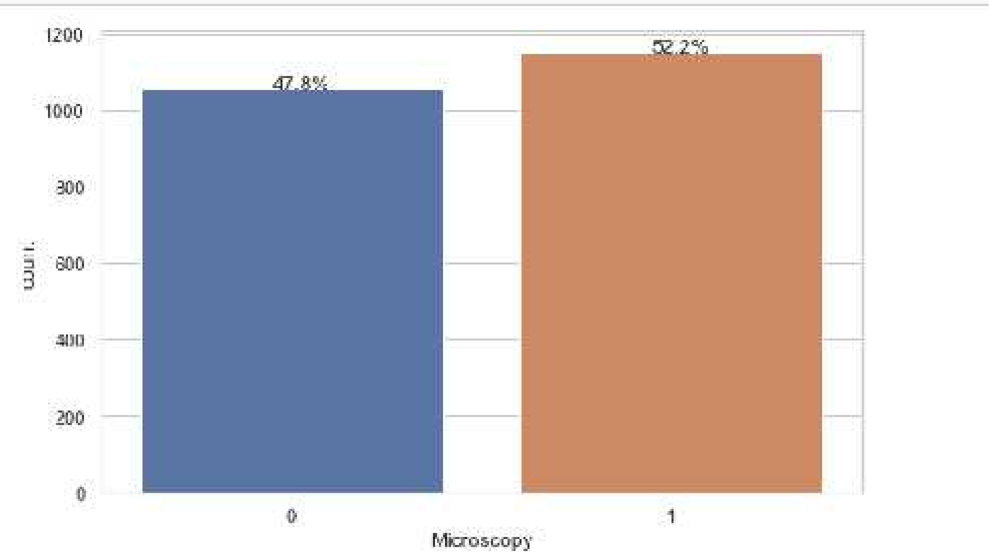
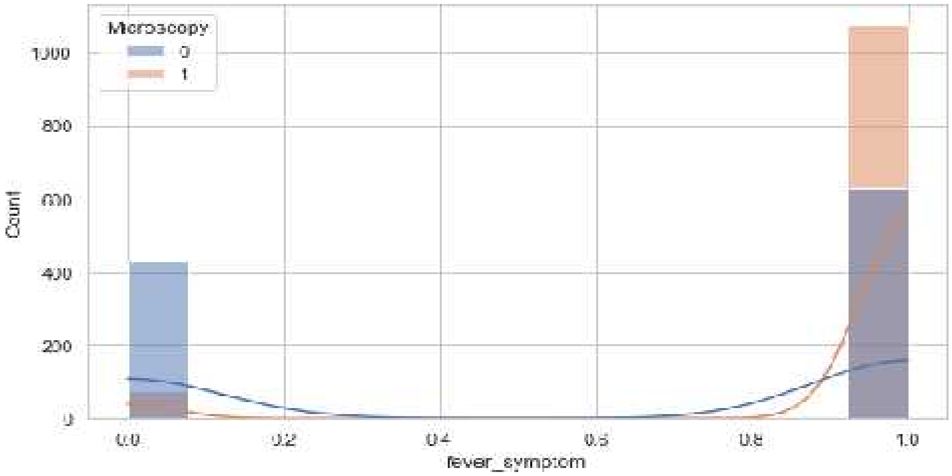


Figure4:2showstheillustrationofaunivariatedescriptionofMicroscopyasatarget.Theanalysis shows52.2%forpositivecasesand47.8%fornegative cases.

4.1.2.2 BivariateAnalysis

Figure4:3

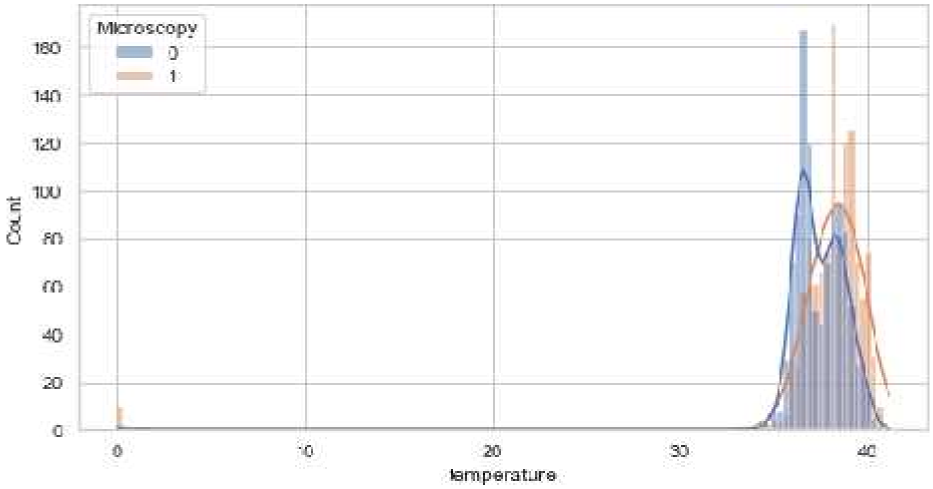
BivariaterelationshipbetweenFeversymptomsandmicroscopy



The image representation in Figure 4:3 shows the relationship between fever symptoms andmicroscopy in a bivariate analysis. The result shows high chances of the patient with malariawhen symptomatic to fever. Similarly, Figure 4:4 shows the relationship between temperatureasa symptomwiththemicroscopy,givinghighchancesofamalariapatient.

Figure4:4

BivariaterelationshipbetweenTemperatureandmicroscopy



4.2 DataPre-processing:

4.2.1 DataEncoding

Thecollectionofdatahasbeengeneratedforthemachine learningapplicationsinthissection. Primarily the set of data is separated into the desired variable, "Microscopy,"and the variables that predict it (X dataset). The ’pd.get_dummies’ method is used toapplyone-hotencodingforprocessingcategoricalfeaturesinsidethepredictorvariables. By converting categorical features to binary features using this encodingtechnique, multicollinearity problems are prevented by making sure that only (n-1)binary columns are formed for each of the n categories. The predicted variables arenow present in the dataset, designated "data2," in a numerical form appropriate formachinelearningtechniques.

4.2.2 Scaling

Wehavecompletedcrucialdatapre-processingoperationsinthissectionsampletogeta dataset ready for machine learning. First, the dataset (data2) was subjected to theMin-MaxScaler,whichconvertsthecharacteristicstoastandardrange,oftenbetween0and1.Scalingisimportantconsideringitguaranteesthatallvariablesimpact machinelearningalgorithmsequally,preventinganyonevariablefromoverridingoverallothers duetodisparitiesintheirscales.

4.2.3 Datasplitting

After scaling, we used the train_test_split function to divide the dataset into training andvalidation sets. Because of this separation, we can train machine learning algorithms on asubset of the data (x_train and y_train) and test their effectiveness on new data (x_test andy_test). The validation set comprises of 729 samples with the same 20 features as the trainingset, which has 1478 samples in total. In order to evaluate the generalization of the approach,the test_size parameter is set to 0.33, which indicates that approximately thirty percent of thedata is putasideforvalidity[80].

4.2.4 BuildingBaseModels

In this segment, welayedthefoundationforbuildingfundamentalmachine-learningalgorithmsinPythonusingthesci-kit-learnmodule.Inclassificationwork,wheretheobjectiveistoclassifydatainto presentcategories,such modelsmaybe employed.

Logistic regression is the primary classifier that is imported. This classifier is frequentlyemployed for classification in binary form, where there are two distinct groups for the desiredvariable.Itusesalinearmixtureofinputattributestosimulatethelikelihoodthatacasebelongsto a particular class. A nice place to start is with logistic regression, which is a fundamentalapproachforclassificationissues.

The GaussianNB classifier is on the technique of Naive Bayes. For assignments combiningcontinuous attributes and classification of text, this classifier is quite helpful. It makes theoften-simplisticpremisethattheattributesareindependent, butinpractice,itcanbeeffective,especiallyfortextdata.

Adifferent approachpotent classifierinthefamilyoflearningensemblesistheRandomForestClassifier. To increase accuracy in prediction and decrease overfitting, it integrates manydecisiontreeclassifiers.Randomforestsarerenownedfortheirdependabilityandversatilityinhandlingdifferent datatypes.Whenestimatingfeatureimportance,theyareespeciallyhelpful.Finally, a flexible classifier that can be applied to projects involving regression as well asclassification is the Decision Tree Classifier. Using feature criteria to iteratively partition thedata, decision trees provide forecasts that consider the dominant class inside each partition.Despite being vulnerable to overfitting, decision trees constitute the foundation for

4.2.5 MakingClassifierModelsInstantiated:

- **lr_mod:**WeestablishedaLogisticRegressionclassifierclass.Thebinaryclassificationdifficultiesareidealforlogistic regression.
- **gb_mod:** Applying the Naive Bayes technique, we build a model of the GaussianNBclassifier. It is preferred because it is straightforward and efficient with specific kindsofdata.
- **rf_mod:** We build a Random Forest Classifier level, a collaborative approach thatcombines different decision trees. Random forests are renowned for accuracy andtoughness.
- **dt_mod:**TheDecisionTreeClassifierinstanceiscreated.Bothregressionandclassification problems requiredecisiontrees.

4.2.6 TrainingandFitting:

Wetrainedeachalgorithmforeveryclassifierusingthefittechniqueusingthesuppliedtrainingdata (x_train and y_train). In order for the models to produce accurate forecasts usingthetraining data, this stage entails determining the ideal parameters. The classifiers (lr_mod,gb_mod, rf_mod, and dt_mod) have been trained to make forecasts on fresh, unforeseen datafollowingtraining.

4.2.6.1 CreatingPredictions:

Four alternative classifier algorithms are used to make forecasts: decision tree (dt_mod),logistic regression (lr_mod), and Gaussian naive Bayes (gb_mod). Every trained algorithm isthenusedtoforecastthetargetattributeonatestdataset(x_test).

The linear equation is used in logistic regression (lr_mod) to produce predictions using theinputfeatures.

GaussianNaiveBayes(gb_mod)estimatestheclassprobabilitybyapplyingstatisticalmethodsandconsideringthateveryattributehasanevendistributionacrosseveryclass.

Toincreaseaccuracyanddecreaseoverfitting,randomforest (rf_mod)mixespredictionsfromvarious decisiontrees,eachtrainedonaselectedportionofthedata.

Recursivelydividingtheinputusingitsfeaturevaluestill itgetstotheleafnodeindicatingtheprojectedclass,thedecisiontree(dt_mod)producespredictions.

4.2.7 ModelEvaluation:

By generating the accuracy scores for four different machine learning classifiers—logisticregression (lr_mod), Gaussian Naive Bayes (gb_mod), random forest (rf_mod), and decisiontree (dt_mod)—the provided method section executes model evaluation. A typical metric forassessing classification models is accuracy, which counts the proportion of cases in the testdataset(x_test)thatpredictedaccuratelyoutofallofthem.

Hereisathoroughdescriptionoftheevaluationprocedure:

- **(lr_mod)LogisticRegression**

Thescoringtechnique,whichcontraststhe model'sforecasts(pred_lr)tothetruetarget values(y_test)inthe testdataset,isused todetermine theaccuracyscoreforlogistic regression.

Thelogisticregressionmodelhasanaccuracyscoreofroughly0.914,meaningthatitcorrectlypredicts theclasslabelsforroughly91.4%ofthecasesinthedata.

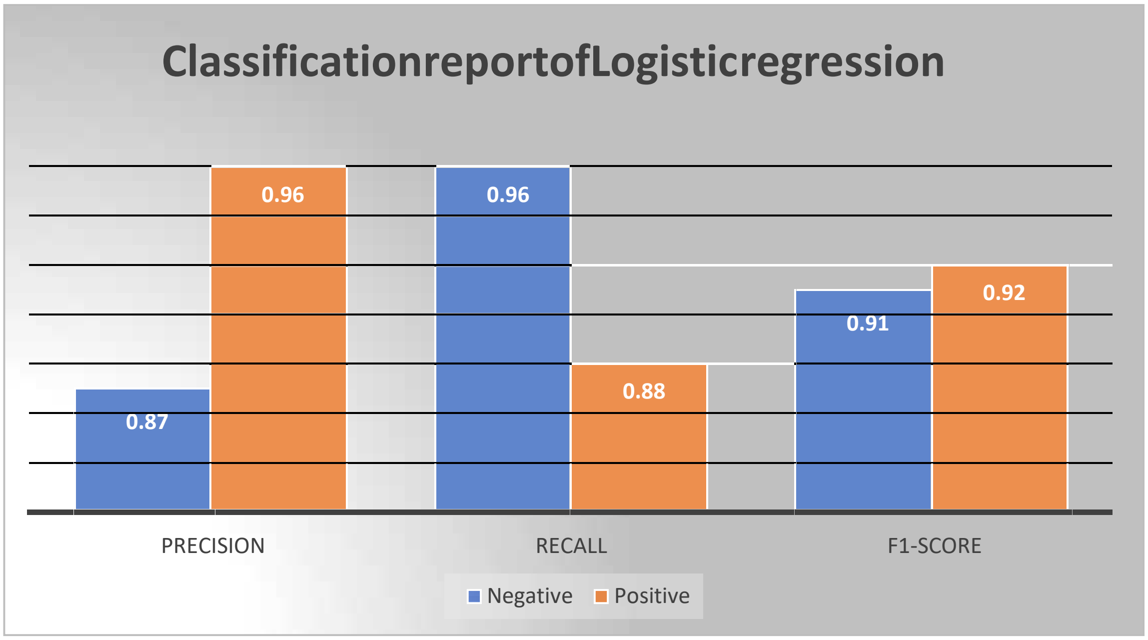
Table4:1:*Classificationreportforlogisticregressionformalariaparasite*

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.87	0.96	0.91	334
Positive	0.96	0.88	0.92	395

As shown in Table 4:1 and Figure 4:5, the logistic regression method's efficacy on a problemwithbinaryclassificationisassessed intheclassificationreport.Themodelshowsgoodoverallpredicting ability with an accuracy rating of about 91.4%. For two classes, "0" and "1,"precision, recall, andF1-scores are given to show how effectively the model accuratelyidentifiesexamplesforeachclass.Precision,recall,andF1-scoreareroughly0.87,0.96,and 0.91 for class "0," accordingly. These metrics are around 0.96, 0.88, and 0.92 for class "1."These findings imply thatthe model successfully distinguishes between the two groups,strikinganappropriateequilibriumbetweenprecisionandrecall.

Figure4:5

Classificationreportforlogisticregressionformalariaparasite



- **NaiveGaussian Bayes(gb_mod):**

The accuracy score for Gaussian Naive Bayes is computed using the scoring technique, samelikeforlogisticregression.Itcontraststhealgorithm'sforecasts(pred_gb)withthetestdataset'strue target values (y_test). Gaussian Naive Bayes' accuracy score is roughly 0.892, whichmeans thatonthetestdataset,thismodelhas anaccuracyofabout89.2%.

Table4:2:Naive*Bayesclassification reportformalaria parasite*

	Precision	**Recall**	**F1-score**	**Support**
Negative	84%	94%	89%	334
Positive	94%	85%	89%	395

AsshowninTable4:2,theclassificationstudyevaluatesthenaivebayesalgorithmefficacyona binary classification work. The algorithm's total success in generating accurate forecasts isindicatedbyitsaccuracyscore,whichisaround89.2%.Thereisprecision,recall,andF1scoresavailablefortheclasses"0"and"1."Precision,recall,andF1-scorearerroughly0.84,0.94,and

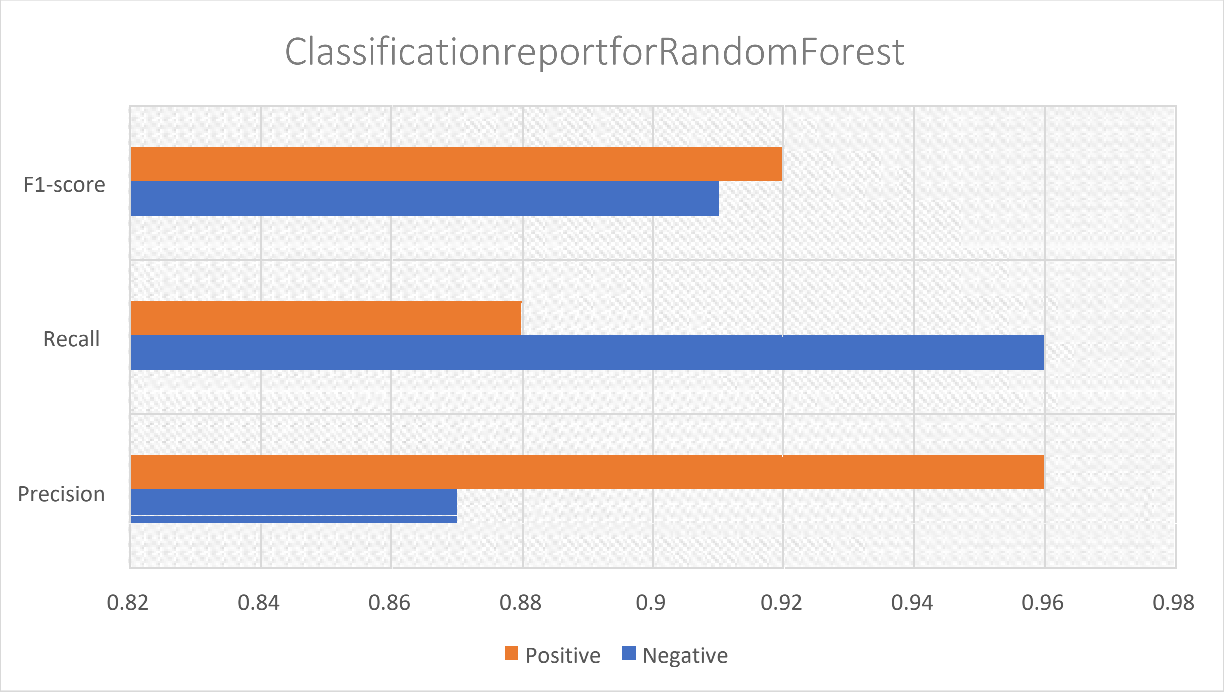
0.89 for class "0," accordingly. These metrics are around 0.94, 0.85, and 0.89 for class "1."These findings show that the Naive Bayes model successfully classifies both groups with adecentmixofrecallandprecision.

- **RF_MOD(RandomForest):**

Similar to the earlier models, the random forest classifier's accuracy score is calculated bycomparing the model's predictions (pred_rf) to the actual target values (y_test). The randomforestgroupalgorithmperformsquitewell,properlycategorizingabout93.3%oftheexamples,accordingtoanaccuracyscoreofroughly0.933asshowninFigure4:6.

Figure4:6

RandomForest reportformalariaparasite



In a binary classification project, the Random Forest Classifier algorithm's effectiveness isevaluated in the classification output. Roughly 93.3% of the forecasts made by the algorithmareaccurate,demonstrating itsgeneralefficacyindoingso.Thereportoffersprecision,recall,and F1 scores for the two groups, denoted by the letters "0" and "1." With regard to class "0,"themodelobtainsprecision,recall,andanF1-scoreofroughly0.84,0.94,and0.89,respectively (see in Figure 4:6). These metrics are around 0.94, 0.85, and 0.89 for class "1"(see in Figure 4:6). These findings show that the random forest classifier approach, which iscomparable to the naive Bayes algorithm for precision and recall, works well in categorizingbothclasses.

- **DecisionsTreeClassifier(dt_mod):**

The decision tree classifier’s scores for accuracy are determined similarly by contrasting itsforecasts (pred_dt) with the real targets (y_test). The above algorithm obtains an accuracy ofroughly 88.9% on the test dataset, according to the decision tree’s accuracy evaluation, whichisroughly0.889as showninFigure4:7.

Figure4:7

Classification reportforDecisionTreeformalariaparasite

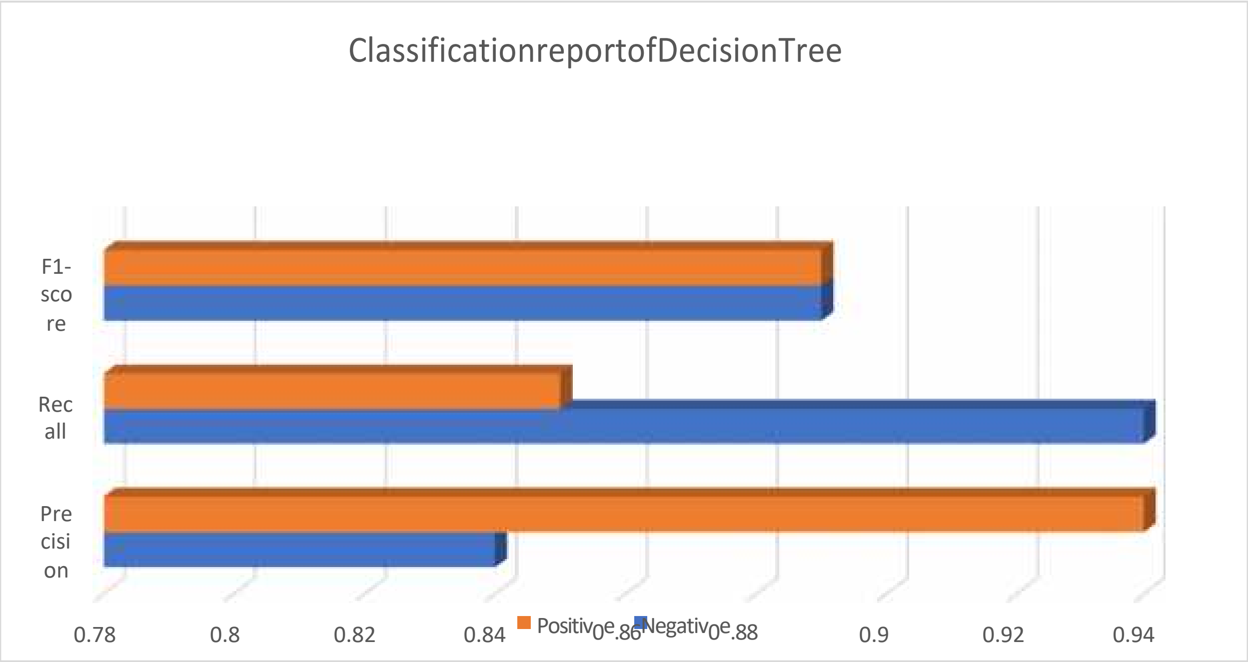


Table4:3:MachinelearningmodelsClassificationreport formalariaparasite

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
Linear regression	0.91	0.92	0.91	0.91
Naivebayes	0.89	0.90	0.89	0.89
Randomforest	0.93	0.90	0.89	0.89

Decisiontree	0.88	0.90	0.89	0.89

Figure4:8

Classificationreportforallmodelsformalariaparasite

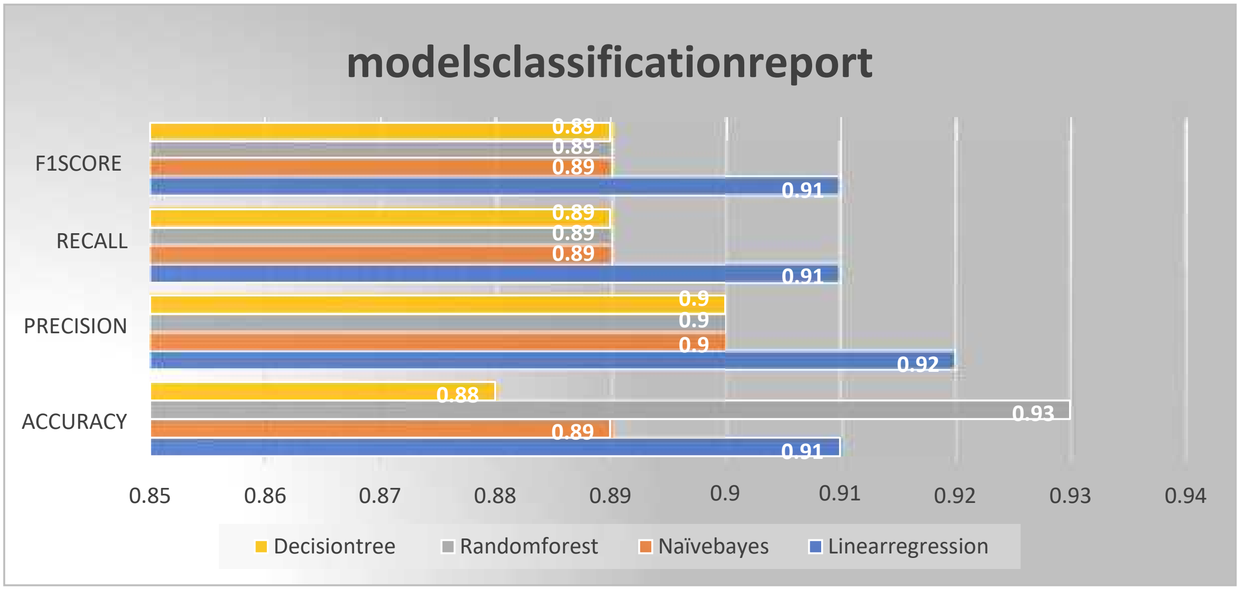


Table4:3showstheresultsfortheclassificationmachinelearningmodels.Whereas,Figure4:8illustratestheclassificationreportofallthemachinelearningmodels.Theclassificationreportprovidesacomprehensiveexaminationofthefouralgorithmsusedtosolvethebinaryclassification problem. Decision Tree, Naive Bayes, Random Forest, and Linear Regression.With the greatest accuracy of 91% and balanced precision, recall, and F1 scores of 0.92, 0.91,and0.91,respectively,linearregressionshowedthebestpredictiveperformance. Whilemaintainingaprecision,recall,andF1scorewell-balancedat0.90,0.89,and0.89,respectively,NaveBayesdemonstratedalittleloweraccuracyof89%. Random Forest surpassed the other models with an accuracy of 93%, precision, recall, and F1scoresof0.90,0.89,and0.89,respectively,provingitsreliabilityinmakingaccuratepredictions.Nottomention,DecisionTreesurpassedNaiveBayesandRandomForest,earningan accuracy of 88% and displaying competitive performance with precision, recall, and F1scoresof0.90,0.89,and0.89,respectivelyas showninFigure4:8.

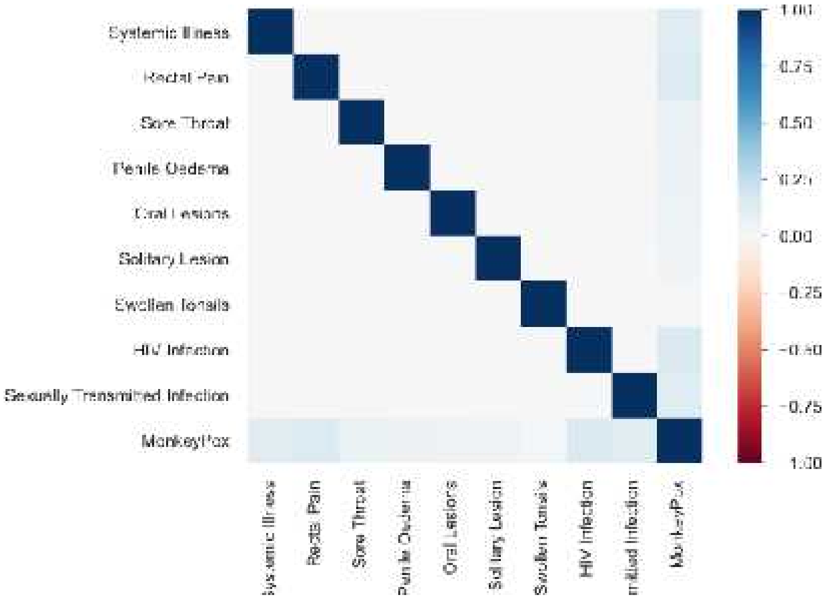
4.3 PredictiveModelsforMonkeyPox:

4.3.1 Variables

Ten criteria plus a dataset with 25,000 items make up this study. Systemic illness, rectal pain,sore throat, penile edema, oral lesions, solitary lesions, swollen tonsils, HIV infection, andsexually transmitted infections are among these characteristics. The "MonkeyPox" targetvariable is the main point of attention. In conclusion, the study examines the connections andtrends among these factors in a dataset with 25,000 observations using machine learning toolsasshowninFigure4:9.

Figure 4:9

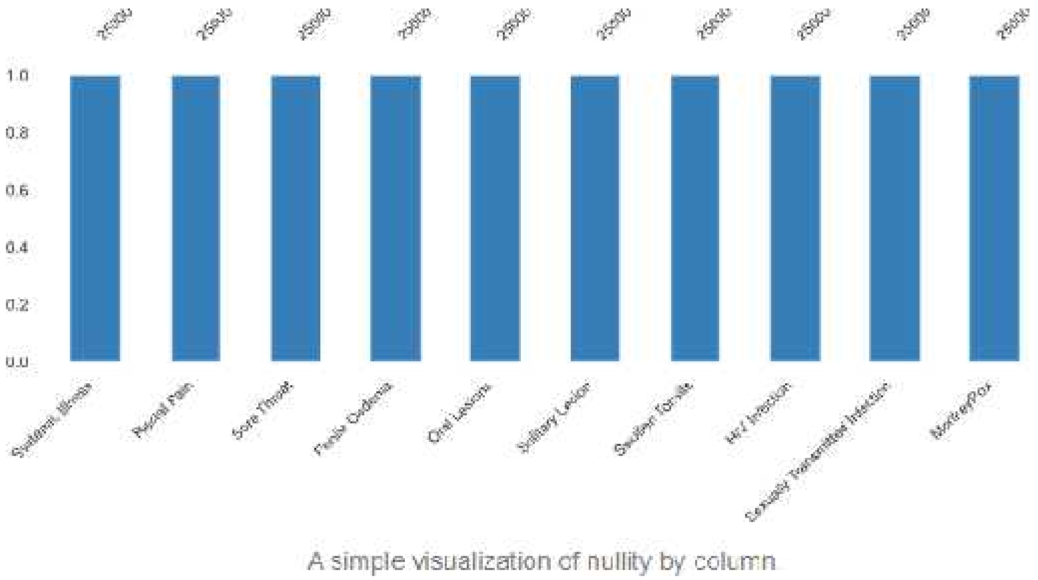
CorrelationHeatmap



As shown in Figure 4:10, visualizes the nullity of each columns. The columns showed no nullfigures.

Figure4:10

Asimplevisualization ofnullitybycolumn.

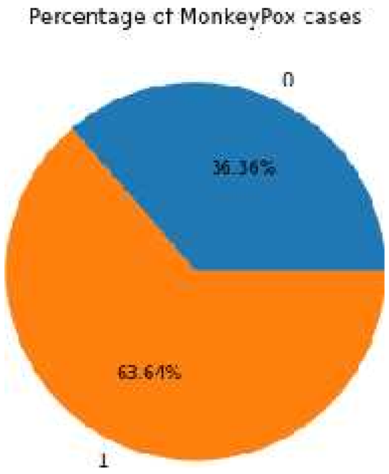


4.3.2 ExploratoryDataAnalysis

4.3.2.1 UnivariateAnalysis

Figure4:11

PercentagecasesofMonkeypox

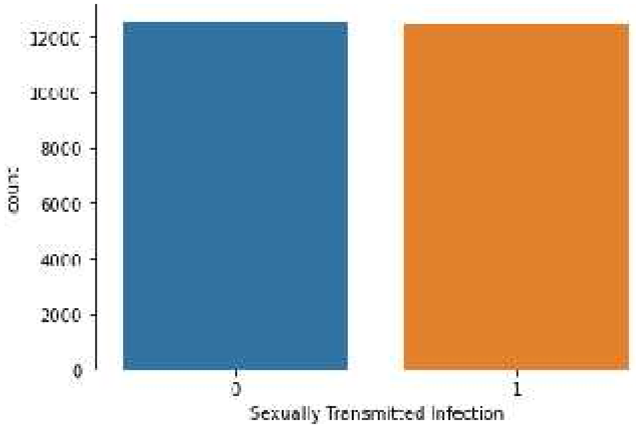


The univariate analysis as illustrated in Figure 4:11, shows the percentage of individuals withnegative and positive results with monkeypox after diagnosis. 64.64% of individuals werecontractedwiththedisease,while36.36%arenegative.

4.3.3 BivariateAnalysis

Figure4:12

BivariaterelationshipofpeoplewithSexuallytransmittedDisease withMonkeyPox.



Theabove Figure4:12 showsthe illustrationoftherelationshipbetweenpeoplewithSexuallytransmitteddiseaseandMonkeyPox.

Figure4:13

BivariaterelationshipofpeoplewithHIVdiagnosedwithMonkeyPox

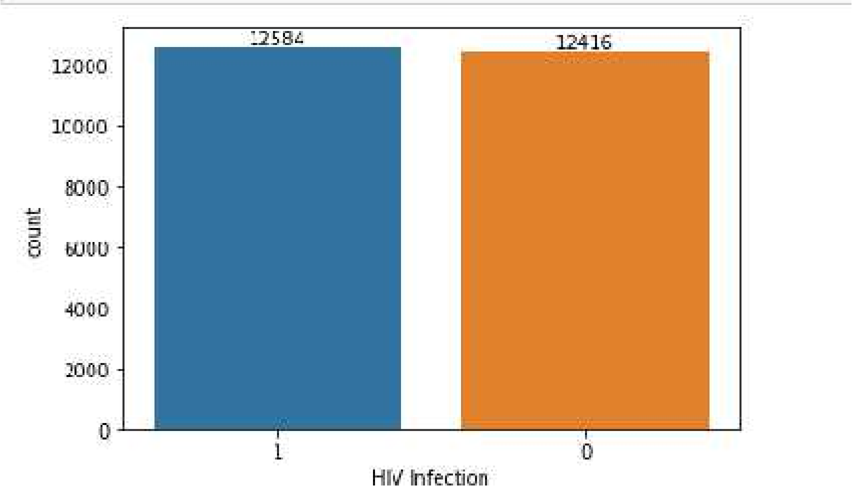
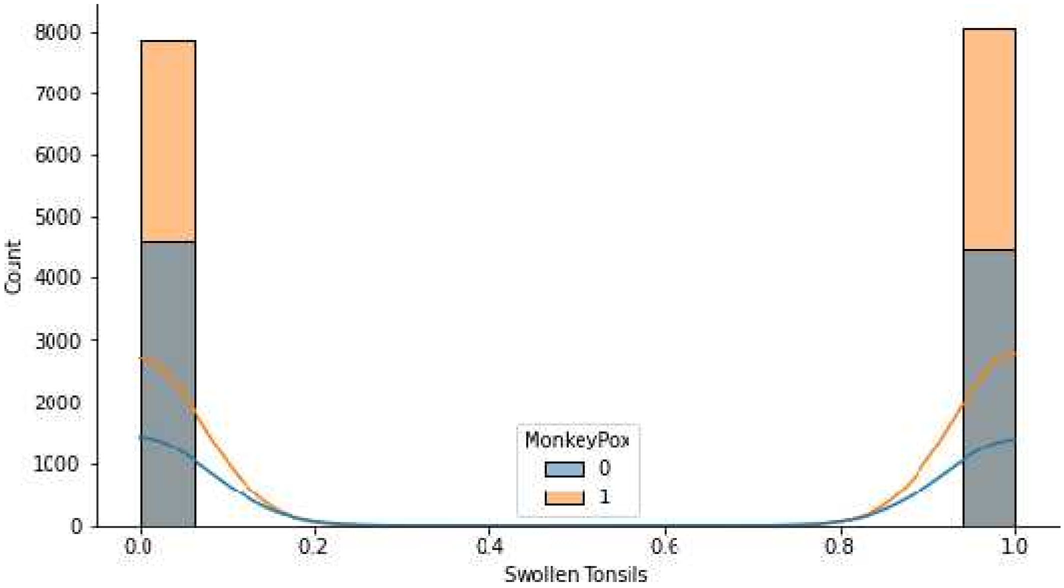


Figure 4:13 represents the bivariate analysis conducted between people diagnosed with HIV,with or without Monkeypox. The data shows that 12,584 positive cases were recorded with12416 negative cases. Similarly, the Figure 4:14 visualizes the patients with swollen tonsilsdiagnose with or without monkeypox. Individuals with monkeypox recorded higher amongdiagnosedindividuals withswollentonsils

Figure4:14

Bivariaterelationshipbetween swollentonsilsandMonkeyPox



4.4 DataPre-processing:

4.4.1 DataEncoding

Operating the MonkeyPox dataset requires segmenting the data into the component of thedataset (X) and the target variable (y). This separation is necessary for the development ofprediction models. In this case, the goal variable is "MonkeyPox," which most likely denoteswhether a person has ever experienced the illness. The feature dataset (X) contains every otherparameter,suchassystemicillness,rectalpain,sorethroat,penileedema,orallesions,solitarylesion, swollen tonsils, HIV infection, sexually transmitted infection, monkeypox. The abilityto forecast or assess whether there is an outbreak of monkeypox using information containedin the attributes is made possible by this data division, making it an essential phase of pre-processing forallthemachinelearning and statisticaltestscarriedout onthe dataset.

4.4.2 Scaling:

To increase the dataset's dependability and effectiveness for analytical and modelling goals, aMinMaxScalerwasdevelopedandused.TheMin-Maxscalerpre-processingtechniqueisusedto alter the dataset, ensuring that all features are scaled to a specific range, frequently between0 and 1. To achieve this, subtract the minimum value from the result and divide it by the rangeof each attribute. By doing this, the dataset as a whole is normalised and the impacts of thedifferentscalesamongtheparametersarediminished.Thisnormalisationiscrucial,especially when utilising machine learning techniques that depend on the size of the input features. Bypreventingbiasesthatcouldarisefromvariationsinthesizeofdistinctfeatures,itensuresthateachfeatureserves the evaluationsandmodelsfairly.

4.4.3 Datasplitting:

Thedatasetwasdividedintotrainingandvalidationsetsusingthe sci-kit-learn'train_test_split' tool. 'x_train' (training features), 'x_test' (validation features), 'y_train'(training target variable), and 'y_test' (validation target variable) were the four subsets thatwere created from the data set. 'scaled_data' (attributes) and 'target' (target parameter) areusefulfor describing thedataset.With'test_size'setto33%,67%ofthedatawillbeusedfortraining and 33% will be retained for validation. The 'random_state' option is set to 42 toensure reproducibility because it fixes the randomly generated seed used in the data splittingprocess. This division is frequently used in machine learning to assess the performance ofmodels onuntesteddataduringthevalidationstage.

4.5 BuildingBaseModels:

Thedatabuiltlaysthefoundationfordevelopingacollectionofmachinelearningmodels,alsoknownasbasemodels.Logisticregression,GaussiannaiveBayes,randomforest,anddecisiontree classifiers are some of the models used in these scikit-learn implementations. Each modelis created for a certain kind of classification task, offering a variety of methods for furtherresearch.Thisestablishesthegroundworkforadditionalmachinelearningprocedureslikedatapre-processing,training,evaluation,andfuture modelperformanceoptimization.

4.5.1 MakingClassifierModelsInstantiated:

Classificationalgorithmsaredevelopedutilizingthetrainingdatathatisreadilyavailable.Themodels are gb_mod (for modified), random forest classifier, decision tree classifier, Gaussiannaive Bayes, and logistic regression. The instantiation uses the relevant classes from Scikit-Learn. Then, every model is fitted to the training data ('x_train' and 'y_train') using the 'fit'approach. This strategy involves changing the internal parameters of every model to produceprecise predictions depending on the supplied features and target variable. When this code isproperly run, trained models are produced that can be used to make predictions using brand-new,unanticipateddata.

4.5.2 CreatingPredictions:

This section focuses on making predictions utilizing the classification algorithms that havealready been trained. The 'predict' technique is used to analyze the test data ('x_test') aftertrainingtheLogisticRegression('lr_mod'),Gaussiannaivebayes('gb_mod'),randomforest classifier ('rf_mod'), and decision tree classifier ('dt_mod') on the training data. Predictedvalues('pred_lr','pred_gb','pred_rf',and'pred_dt')areproducedforeachmodelasaconsequence. Given the characteristics of the test dataset, these predictions represent theexpectedresultsofthemodels.Theeffectivenessandprecisionofthemodelscanbeevaluatedby comparing these predictions to the actual target values ('y_test'). The predictions play acritical role in determining how well each model generalises to new data and in assessing themodels'overallsuccessinaddressingthegivencategorizationchallenge.

4.6 ModelEvaluation:

On the specified test dataset ('x_test', 'y_test'), each of the four models (logistic regression,naive Bayes, random forest classifier, and decision tree classifier) had accuracy scores of 1.0,indicatingaflawlessclassification.A100% accuracyratingindicatesthatthemodelscorrectlyclassified each incident in the test set. While high accuracy is desired, it's important to exercisecaution because perfect precision could be a sign of overfitting, particularly if the algorithmswerecreatedontheexactsamevariables asthe testing.

To guarantee resilience and generalization to new data, it is advised to further assess themodel's efficiency using more metrics including accuracy, recall, and F1-score as well as tousecross-validationapproaches.Additionally,it'simportanttoevaluatemodelsusingavarietyofdatasets andlookatpossibleproblems suchasdata leaking.

- **(lr_mod)LogisticRegression**

The test dataset ('x_test', 'y_test') was perfectly classified according to the logistic regressionapproach, which received a desirable accuracy score of 1.0. All class's recall, precision, andF1-score (0 and 1) are thoroughly broken down in the classification report. Precision, recall,and F1-score for both classes are all provided as 1.0, signifying flawless performance. Thesupport column displays how many instances there are of each class. The logistic regressionmodel outperformed all other models in this test, as seen by its precision, recall, and F1-scoremacroandweightedaveragesbeing all1.0as shown inTable 4:4.

Table4:4:ClassificationreportforLogisticregression

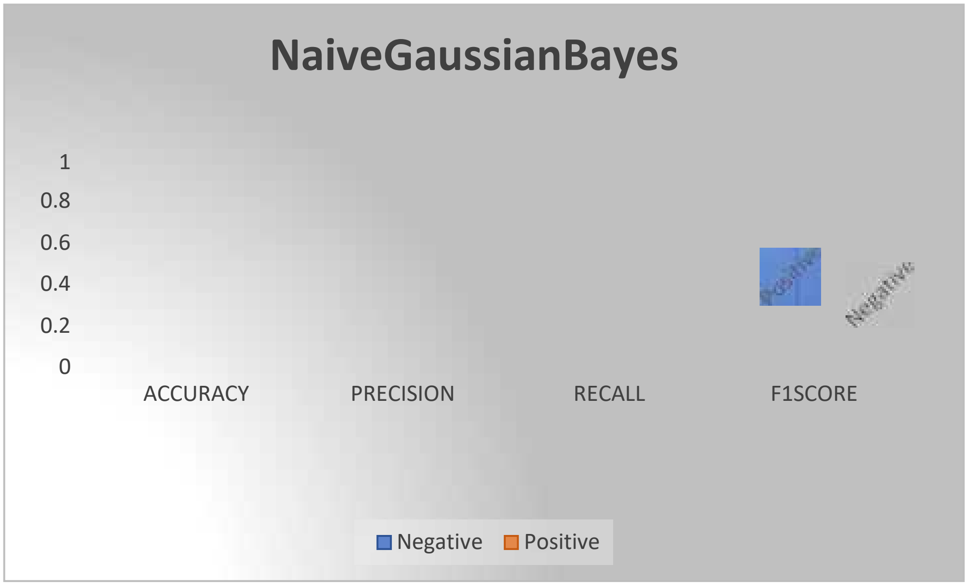
	ACCURACY	PRECISION	RECALL	F1 SCORE
Negative	1.0	1.0	1.0	1.0
Positive	1.0	1.0	1.0	1.0

- **NaiveGaussianBayes(gb_mod):**

TheNaiveBayesalgorithmlikewiseperformedverywell,gettingaperfectaccuracyscoreof 1.0 on the test dataset ('x_test', 'y_test'). The model's accuracy is additionally broken into fourcategories: precision, recall, and F1-score for each class (0 and 1) in the classification report,every single one of which has been identified as being 1.0 as shown in Figure 4:15. Thisrepresentsflawlessclassificationforbothcourses.

Figure4:15

ClassificationreportforNaiveBayes



Thegeneralizedandweightedmeanofprecision,recall,andF1-scoreareallrecordedasbeing1.0,indicating thattheNaiveBayesalgorithmfaredprettywellconsistentlyinthistest.Extra testing on several datasets with consideration for potential overfitting is necessary to ensurethemodel's accuracyindifferentcontexts.

- **RandomForestClassifier:**

A flawless accuracy score of 100% on the test dataset ('x_test', 'y_test') was also achieved bytheRandomForestClassifier,furtherdemonstratingitsexceptionalperformance.Theclassificationreportgivesathoroughbreakdownoftheprecision,recall,andF1-scoreforeachclass (0 and 1), all of which are given as 1.0, demonstrating perfect categorization across allclasses.Theinstancesofeveryclassareshowninthe supportcolumn.

Table4:5:ClassificationreportforLogisticregression

	PRECISION	ACCURACY	RECALL	F1 SCORE
Negative	1.0	1.0	1.0	1.0
Positive	1.0	1.0	1.0	1.0

Furthermore, the precision, recall, and F1-score macro and weighted averages are all given as100%, highlighting the Random Forest Classifier's exceptional overall performance in thisparticular evaluation. Testing the algorithm on various dataset is essential to determining itscapacitytogeneralizeandanypossibilityofoverfittingasshowninTable4:5.

- **DecisionTreeClassifier:**

Every instancein the testsetwas accurately classified,accordingtothe Decision TreeClassifier's classification report, which demonstrates remarkable performance and has anaccuracy score of 1.0. Precision, recall, and F1-score should all be set to 1.0 for a morethorough study. This shows that the model, as shown by the F1-score, has an outstandingbalance between precision and recall and is exceptionally accurate in correctly identifyinginstances of monkeypox. The Decision Tree Classifier displays impressive generalization andclassificationperformanceonthe MonkeyPoxdatasetasshowninFigure 4:16.

Figure4:16

DecisionTreeClassifier

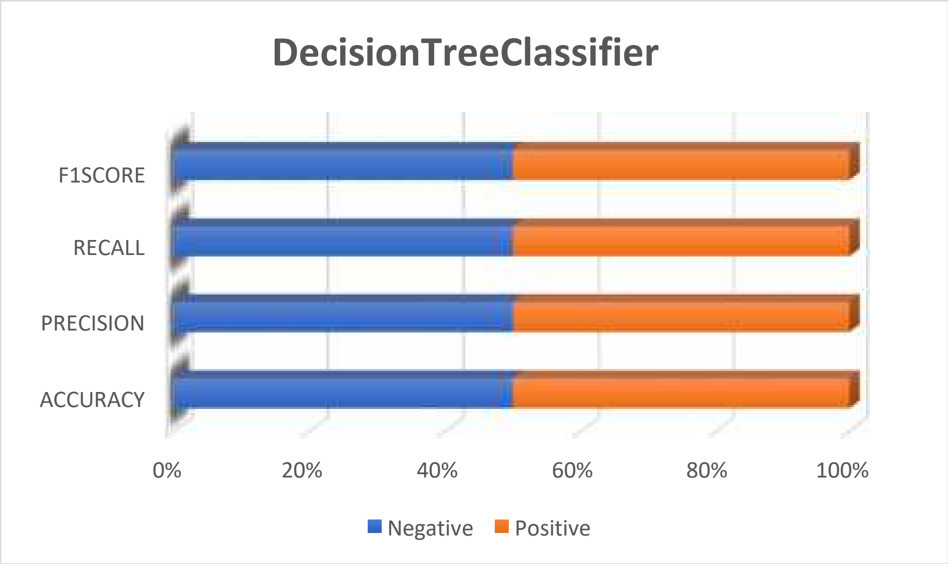


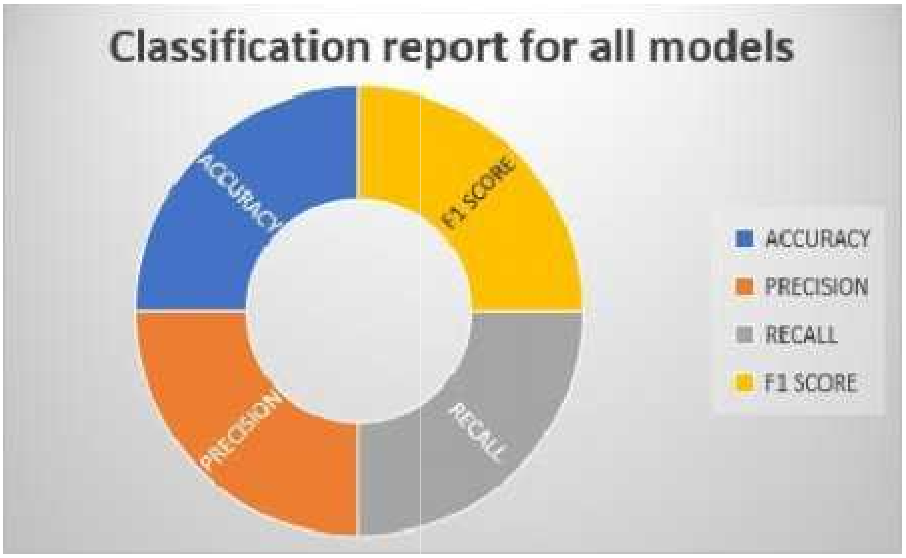
Table4:6:Classificationreportforthemodels

MODEL	ACCURACY	RECALL	PRECISION	F1 SCORE
Linearregression	1.00	1.00	1.00	1.00
Naivebayes	1.00	1.00	1.00	1.00
Randomforest	1.00	1.00	1.00	1.00
Decisiontree	1.00	1.00	1.00	1.00

The evaluation results for the classification algorithms Decision Tree Classifier, RandomForest Classifier, Naïve Bayes, and Logistic Regression on the MonkeyPox dataset showexceptional performance. The accuracy score for the test set was 1.0 since every case in thetest set was correctly classified for every model. The precision, recall, and F1-score measuresshowthatthemodelsarerobust,withvaluesof1.0forbothclasses(0and1).Accordingtothe F1-score,themodelsappeartoexhibitasatisfyingmixofprecisionandmemory,highaccuracyincorrectly recognizingMonkeyPoxcases,andexcellentrecallingatheringallrealMonkeyPoxcases as showninTable4:6.

Figure4:17

Visualization forClassificationreportforthemodels



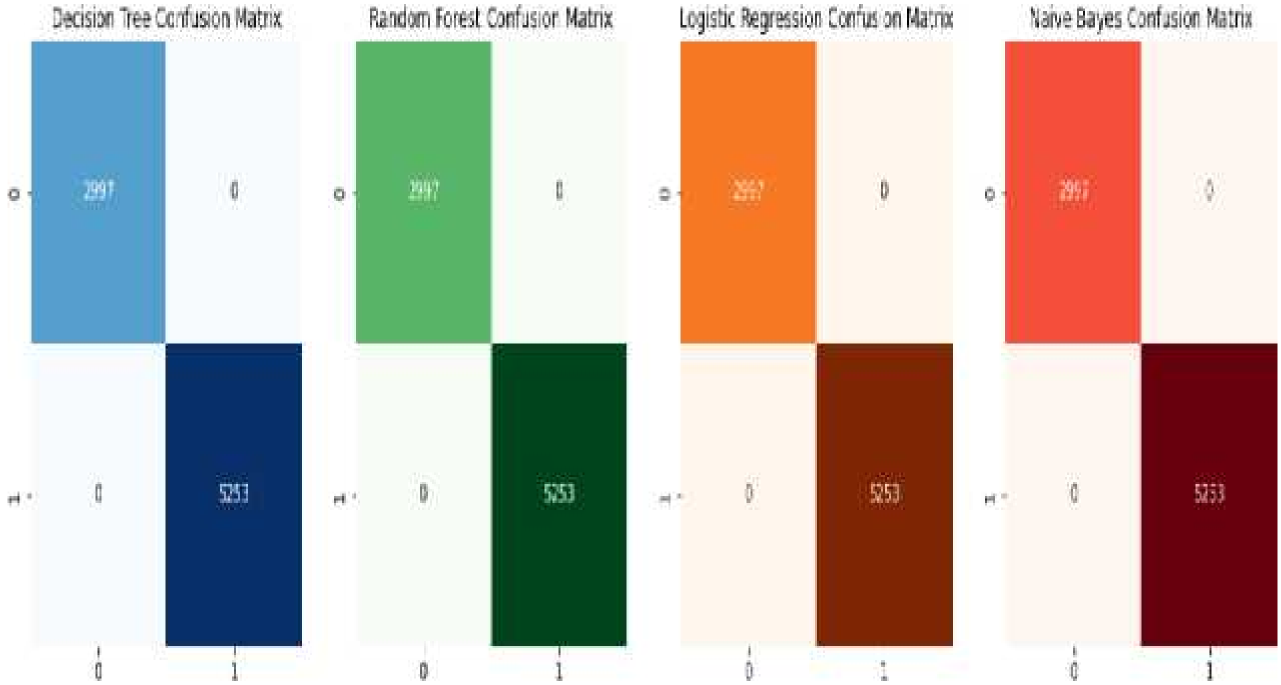
Inconclusion,allmodelsdemonstrateoutstandinggeneralisationandpredictionaccuracyusingthe Monkey Pox dataset. The models consistently and reliably provide high scores across arange of metrics, suggesting that they are well suited for the task of classifying instances ofmonkeypoxasvisualizedinFigure4:17.

- **ConfusionMatrix:**

The non-existence of any kind of off-diagonal (misclassification) components demonstratesthe excellent classification performance displayed by the confusion matrices across all themodels (Decision Tree, Random Forest, Logistic Regression, and Naive Bayes). A squarematrix keeps track of the number of forecasts categorized as true positive (TP), true negative(TN),falsepositive(FP),andfalsenegative (FN)asshown inFigure 4:18.

Figure4:18

Confusionmatrixresult



CHAPTER5

Discussion

A common method of analysis was used in both studies on monkeypox and malaria to usemachinelearningtopredictinfectiousdiseases.Theanalysessharesimilaritiesindatacollecting,pre-processing,andmodelconstructionwhilehavinguniquediseasesbroughtonbyvarioussources.Inbothresearch,variablescomprisingclinicalsymptoms,resultsofdiagnostictests, and pertinent health indicators are examined. In-depth pre-processing ensures uniformandobjectivesourcesforthemachinelearningalgorithmsbyusingmethodslikedataencodingand scaling. The models used throughout both investigations include Logistic Regression,GaussianNaiveBayes,RandomForest,andDecisionTree classifiers.

Theresultsonmalariademonstratestrongmodelperformance,withlogisticregressionyieldingan accuracy of about 91%. A reasonable classification for both malaria-positive and negativeresults can be seen in the precision, recall, and F1 scores. The choice of the most appropriatemodel for malaria prediction is aided by a detailed evaluation that highlights the relativeadvantagesanddisadvantages ofeachclassifier.

Thestudygetsanoutstandingperformanceratingof100%fortheclassificationofmonkeypoxusingLogisticRegression,NaiveBayes,RandomForest,andDecisionTreeclassifiers.Precision, recall, and F1-scores regularly exceed 1.0 for both positive and negative situations,demonstratingthemodels'dependability.Theconfusionmatriceshighlighttheperfectclassificationperformanceofallmodels,withoutanyincidentsofmisclassificationbeingseen.Thisimpliesthatmachinelearningalgorithmsarereliableandresilientinpredictingmonkeypoxbasedonclinicalandsymptomaticdata.

TheUnivariateDescriptionofMicroscopyisdisplayed inFigure4:2,withanoverallincidenceof52.2%casesthatwerepositiveand47.8%negativecases,thisimageprovidesacrystal-clearunivariate representation of malaria cases. The foundation for additional analysis is formed bythisvisualdepiction,whichgivesaquickcomprehensionoftheclassdistributionofthedataset.The Bivariate AnalysisofFeverSymptoms andMicroscopyisdisplayedin Figure4:3.

This graph examines the connection between malaria symptoms and fever symptoms andshows a big link. The visual evidence demonstrates the value of fever as a symptom inidentifying malaria cases, which is important knowledge for both medical diagnosis and thedevelopmentofmachinelearningmodels.

The connection between temperature and microscopy is seen in Figure 4:4 as explored in thegraph. The visualization emphasizes the link between high temperatures and incidences ofmalaria,providingusefulinformationfrombothamedicalandpredictivemodellingstandpoint.

Theclassificationreportforlogisticregressionisshowninthetable.TheLogisticRegressionmodel’s performance indicators are broken down in great detail in Table 4:1. For all classes (0and 1), accuracy, precision, recall, and F1-score are shown, providing a thorough assessmentofthemodel’s predictionskills.

The random forest classification report is displayed in Figure 4:6. The performance indicatorsfor the random forest classifier are summarised in this graphic. With a focus on both positiveand negative situations, the visual representation enables an instant assessment of the model’sefficiency,precision,recall,andF1-score.

In a similar vein, Figure 4:11 displays the univariate description of monkeypox cases formonkey pox prediction. A graphic analysis of monkeypox cases revealed a predominance of64.64%positivecasesand36.36%negativecases.Thecontextforfurtheranalysisisestablishedbythis understandingoftheclassdistribution.

A bivariate analysis of HIV and monkeypox is also shown in Figure 4:13. The link betweenHIVcasesandmonkeypoxcasesisexamined inthefigure,whichshedsinformationontheco-occurrence of these diseases. The graphic representation makes it easier to comprehend howmanyhealthindicators interactwithoneanother.

Table 4:5 also includes the logistic regression classification report. This table provides acomprehensiveclassificationreportforthelogisticregressionmodelinthecontextofmonkeypox, similar to the study on malaria. The measurements provide a thorough evaluationofthealgorithm’sprecisionanddependability.thevisualisationfortheclassificationreportforthemodelsis showninFigure4:17.

The classification reports for all models—logistic regression, naive Bayes, random forest, anddecisiontree—canbeshowntogetherinthispicture.Arapidassessmentofmodelperformances acrossvariousmetricsis possiblewiththe graphicalsummary.

Thetwostudiesemployedunivariateandbivariatevisualizationstoprovideinformationonthenumberofcases ofdiseases andtheconnectionsbetweenkeymedicalindicators.

A nuanced knowledge of each class’s accuracy, precision, recall, and F1-score are possiblebecausetothe extensivebreakdownofmodelperformance dataprovidedbythe classification reportsinthetables.Theflawlessaccuracyscoresinbothexperimentsgeneratequestionsregardingpossibleoverfitting,whichcallsforacautiousreadingofthefindings. Althoughnotexpresslystated,theconfusionmatricesareassumedtobefreeofmisclassifications,supportingthe excellentperformanceofthe models. Thisstudyisdistinctivelyoutstandingwhencompared tostudiesby[44],[45],[46],[47],and [48]focusesontwodeadlyinfectiousdiseases,malariaandmonkeypox,whichposeimportantpublic health challenges. This study employs machine learning to predict the prevalence andspread rates of both malaria and monkeypox, whereas earlier studies only employed machinelearning methods to detect, categorize, or analyse a specific disease. For example, Morang'a,Collins M., et al. tested six distinct methods for machine learning before settling on one. Theyemployed a three-layer artificial neural network (ANN) for the multiple classification of UM,SM, and nMI. Furthermore, binary classifiers were developed to identify the characteristicsthatmaydistinguishUMorSMfromnMI.Inasimilarstudy,UzunOzsahin,Dilber,etalfocuson the challenges involved in discerning between chickenpox and monkeypox due to thesimilarityoftheskinlesions,whichmayresultinmisdiagnosisandadditionaldiseasedissemination, urgingdeeplearningtoolssuchastheCNNalgorithm,AlexNet,andVGGNet.

Therefore, none of these studies implored two different datasets of two different diseasesapplying the same machine learning models in the classification. This study is novel in itsapproach.

CHAPTER6

ConclusionandRecommendation

This chapter outlines the enhancements that this study has contributed to the establishedresearchworldinthescientificliteratureaswellastheimplicationsofitsresultsandalsoservesas a summary of the full study. The study’s limitations are then discussed,followed bysuggestionsforadditionalresearch atthechapter’sconclusion.

6.1 Conclusion

In summary, the research on monkeypox and malaria applied machine-learning algorithms toimprove disease detection and forecasting. Univariate and bivariate analyses offer importantinsights into the incidence and connections connecting various risk factors for these illnesses.Thedatasetshavebeen createdforsystem trainingandevaluationbycareful datapre-processing,comprisingencoding,scaling,andsplitting.

Numerous classifiers, including Logistic Regression, Gaussian Naive Bayes, Random Forest,andDecisionTree,areinstantiatedtoshowtheireffectivenessincorrectlypredictingthecourse ofmonkeypoxandmalariaillness.Thesimulationsperformedexceptionallywell.Theaccuracy, precision, recall, and F1-score metrics are highlighted in the classification reports,demonstratingthemodels’dependabilityandresiliencetothetwoailments.

Theconfusionmatriceshighlight themodels’precisionindifferentiatingbetweenpositiveandnegative situations while highlighting the lack of any misclassifications. An in-depth analysisofeachalgorithm’seffectivenessisgivenbythetabularmetricsandthegraphicrepresentationsofcategorizationreports.

Future work on these subjects could focus on overcoming difficulties with generalization,ethical issues, and practical applications. For the models to be useful in various healthcarecontexts,cross-validation strategies,teamwork with medical professionals,andcontinualmodel updates are crucial. Furthermore, extending the capability of addressing multi-classcasesandaddinguniquefeatureswouldhelpmachinelearningtechniquescontinuetoimproveand becomemoreapplicable tothe diagnosisandtreatmentofdiseases.

6.2 Recommendations

The prospects for the future of the research on malaria and monkeypox include a number ofcriticalcomponentstoimprovetheusefulnessanddependabilityofmachinelearningalgorithms in the medical field. It is crucial to ensure generalization over various datasets andpopulationsingeneral,usestrongcross-validationmethods,andcarryoutexternalvalidationson different datasets. For seamless incorporation into healthcare processes, clinical decisionsupport systems must move towards real-time applications, incorporate more features forimproved predictive capability, and place a priority on model interpretability. The ethical andresponsible use of AI in healthcare will also be aided by dynamic model updating, teamworkwith medical personnel, and resolving ethical issues. Also, expanding studies to include novelillnessclassesorseveritylevelsandhandlingmulti-classclassificationsettingswould increasethe reach and significance of the application of machine learning in disease prediction.

Thefindingsprovideastrongplatformforfurtherstudyandencouragetheinvestigationofvariousdatasets and demographics in order to validate and generalize the machine learning models.Future studies may benefit from using cross-validation methods and evaluating models onseveral datasets due to potential overfitting issues with perfect accuracy ratings. The use ofvisualandtabularrepresentations,whichpromoteclarityandcomprehensibilitywhencommunicatingcomplicatedmachinelearningresultstoboth scientificandprofessionalaudiences, is an excellent example of successful machine learning outcomes transmission.Similar methods could also be used in predicting other diseases, using multiple machinelearningtools.

Reference

1. Lozano,J.M.,&Muller,S.(2022).Monkeypox:Potentialvaccinedevelopmentstrategies. *TrendsinPharmacologicalSciences*.
2. Martins-Filho,P.R.,Nicolino,R.R.,& da Silva,K.(2022).Incidence,geographicdistribution, clinical characteristics, and socioeconomic and demographic determinants ofmonkeypox in Brazil: A nationwide population-based ecological study. *Travel MedicineandInfectious Disease*.
3. WorldHealthOrganization(2023)"Mpox(MonkeyPox)https://www.who.int/news-room/fact-sheets/detail/monkeypox
4. Pittman, P. R., Martin, J. W., Kingebeni, P. M., Tamfum, J. J. M., Wan, Q., Reynolds, M.G., ... & Huggins, J. W. (2022). Clinical characterization of human monkeypox infectionsintheDemocraticRepublicoftheCongo. *MedRxiv*,2022-05.
5. D'Antonio,F.,Pagani,G.,Buca,D.,&Khalil,A.(2023).Monkeypoxinfectioninpregnancy:asystematicreviewandmetaanalysis.*Americanjournalofobstetrics&gynecology MFM*,*5*(1),100747.
6. Martínez, J. I., Montalbán, E. G., Bueno, S. J., Martínez, F. M., Juliá, A. N., Díaz, J. S., ...& Arnáez, A. A. (2022). Monkeypox outbreak predominantly affecting men who have sexwithmen,Madrid, Spain,26Aprilto16June2022.*Eurosurveillance*, *27*(27),2200471.
7. Luo,Q.,&Han,J.(2022).Preparednessforamonkeypoxoutbreak.*InfectiousMedicine*.
8. Philpott,D.,Hughes,C.M.,Alroy,K.A.,Kerins,J.L.,Pavlick,J.,Asbel,L.,...&Johnson, S. (2022). Epidemiologic and clinical characteristics of monkeypox cases—United States,May17–July22,2022.*MorbidityandMortalityWeeklyReport*,*71*(32),1018.
9. Shuvo, P. A., Roy, A., Dhawan, M., Chopra, H., & Emran, T. B. (2022). Recent outbreakof monkeypox: Overview of signs, symptoms, preventive measures, and guideline forsupportivemanagement.*InternationalJournalofSurgery(London,England)*,*105*,106877.
10. MonkeyPoximage(2023)https://www.gavi.org/vaccineswork/five-charts-monkeypox-past-andpresent?gclid=CjwKCAjwgZCoBhBnEiwAz35RwnZaMPpgHEF5affVZt94r_7BUdlj-zsOgHXs138si5pKeAA1ZjdmiBoC3nUQAvD_BwE
11. Zachary, K. C., & Shenoy, E. S. (2022). Monkeypox transmission following exposure inhealthcare facilities in nonendemic settings: Low risk butlimitedliterature. *InfectionControl&HospitalEpidemiology*,*43*(7),920-924.

12. Antinori,A.,Mazzotta, V.,Vita,S.,Carletti,F.,Tacconi,D.,Lapini,L.E.,...&Nicastri, E.(2022).Epidemiological,clinicalandvirologicalcharacteristicsoffourcasesofmonkeypoxsupporttransmissionthroughsexualcontact,Italy,May2022.*Eurosurveillance*,*27*(22),2200421.
13. Monkeypox outbreak (2023)https://www.cdc.gov/poxvirus/monkeypox/response/2022/world-map.html
14. Chen,X.,Yuan,W.,Yang,X.,Shi,Y.,Zeng,X.,Huang,J.,...&Li,S.(2023).Ultrasensitive and Specific Identification of Monkeypox Virus Congo Basin and WestAfrican Strains Using a CRISPR/Cas12b-Based Platform. *Microbiology Spectrum*, *11*(2),e04035-22.
15. Mitjà, O., Ogoina, D., Titanji, B. K., Galvan, C., Muyembe, J. J., Marks, M., & Orkin, C.M.(2023).Monkeypox.*The Lancet*,*401*(10370),60-74.
16. mondiale de la Santé, O., & World Health Organization. (2023). Epidemiology of humanmonkeypox (mpox)–worldwide, 2018–2021–Épidémiologie de la variole simienne chezl’humain–danslemonde,2018-2021.*WeeklyEpidemiologicalRecord=Relevéépidémiologiquehebdomadaire*,*98*(03),29-36.
17. mondiale de la Santé, O., & World Health Organization. (2023). Epidemiology of humanmonkeypox (mpox)–worldwide, 2018–2021–Épidémiologie de la variole simienne chezl’humain–danslemonde,2018-2021.*WeeklyEpidemiologicalRecord=Relevéépidémiologiquehebdomadaire*,*98*(03),29-36.
18. McCollum,A.M.,Shelus,V., Hill,A.,Traore, T.,Onoja,B.,Nakazawa,Y., ...&Lewis, R.(2023).Epidemiologyofhumanmpox—Worldwide,2018–2021.*MorbidityandMortalityWeeklyReport*,*72*(3),68.
19. McCollum,A.M.,Shelus,V., Hill,A.,Traore, T.,Onoja,B.,Nakazawa,Y., ...&Lewis, R.(2023).Epidemiologyofhumanmpox—Worldwide,2018–2021.*MorbidityandMortality WeeklyReport*,*72*(3),68.
20. Ogoina,D.(2022).Sexualbehavioursandclinicalcourseofhumanmonkeypoxin Spain. *TheLancet*,*400*(10353),636-637.
21. Hraib, M., Jouni, S., Albitar, M. M., Alaidi, S., & Alshehabi, Z. (2022). The outbreak ofmonkeypox2022:Anoverview.*Annals ofMedicine andSurgery*,*79*,104069.
22. Nakhaie, M., Arefinia, N., Charostad, J., Bashash, D., Haji Abdolvahab, M., & Zarei, M.(2023). Monkeypox virus diagnosis and laboratory testing. *Reviews in Medical Virology*,*33*(1),e2404.

23. Soheili,M.,&Nasseri,S.(2022).Monkeypox:virology,pathophysiology,clinicalcharacteristics, epidemiology, vaccines, diagnosis, and treatments. *Journal of pharmacyandpharmaceuticalsciences*,*25*(25),322-297.
24. Jiang,R.M.,Zheng,Y.J.,Zhou,L.,Feng,L.Z.,Ma,L.,Xu,B.P.,...&Shen,K.L.(2023).Diagnosis, treatment, and prevention of monkeypox in children: an experts' consensusstatement.*WorldJournalofPediatrics*,*19*(3),231-242.
25. Khattak, S., Rauf, M. A., Ali, Y., Yousaf, M. T., Liu, Z., Wu, D. D., & Ji, X. Y. (2023).Themonkeypoxdiagnosis,treatmentsandprevention:Areview.*FrontiersinCellularandInfectionMicrobiology*,*12*,2005.
26. Gupta, A.K., Talukder,M., Rosen,T.,& Piguet, V. (2023).Differential Diagnosis,Prevention,andTreatmentofmpox(Monkeypox):AReviewforDermatologists.*American JournalofClinicalDermatology*,1-16.
27. Pacheco, M. A., & Escalante, A. A. (2023). Origin and diversity of malaria parasites andotherHaemosporida.*Trends inParasitology*.
28. Malaria parasite (2023) https://www.who.int/news-room/fact-sheets/detail/malaria#:~:text=Disease%20burden&text=cases%20in%202020.-,The%20estimated%20number%20of%20malaria%20deaths%20stood%20at%20619%20000,to%20625%20000%20in%202020.
29. Wichers-Misterek,J.S.,Binder,A.M.,Mesén-Ramírez,P.,Dorner,L.P.,Safavi,S.,Fuchs,G.,...&Gilberger,T.W.(2023).AMicrotubule-AssociatedProteinIsEssentialforMalariaParasiteTransmission.*Mbio*,*14*(1),e03318-22.
30. Fornace,K.M.,Laporta,G.Z., Vythilingham,I.,Chua, T.H.,Ahmed,K.,Jeyaprakasam, N.K.,...&Lau,Y.L.(2023).Simianmalaria:anarrativereviewonemergence,epidemiologyandthreattoglobalmalaria elimination.*TheLancetInfectiousDiseases*.

31. WorldHealthOrganization.*WorldMalariaReport2019.Geneva.*;2019.https://www.who.int/publications-detail/world-malaria-report-2019

32. Rubinger,L.,Gazendam,A.,Ekhtiari,S.,&Bhandari,M.(2023).Machinelearningandartificialintelligenceinresearchandhealthcare.*Injury*,*54*,S69-S73.

33. Symptoms of malaria (2023) https://www.who.int/news-room/questions-and-answers/item/malaria?gclid=CjwKCAjw9-6oBhBaEiwAHv1QvGNxvtk3-h-HmZwkyIKOvuwrEDhwl-J1iuIU5-jKdZ8kKJlGPBmdHRoCztoQAvD_BwE

34. Andagalu,B.,Watson,O.J.,Onyango,I.,Opot,B.,Okoth,R.,Chemwor,G.,...&Kamau, E. (2023). Malaria Transmission Dynamics in a High-Transmission Setting of WesternKenyaandtheInadequateTreatmentResponsetoArtemether-LumefantrineinanAsymptomaticPopulation.*ClinicalInfectiousDiseases*,*76*(4),704-712.

35. Koffi, A. A., Camara, S., Ahoua Alou, L. P., Oumbouke, W. A., Wolie, R. Z., Tia, I. Z., ...&N’Guessan,R.(2023).AnophelesvectordistributionandmalariatransmissiondynamicsinGbêkêregion,centralCôted’Ivoire.*MalariaJournal*,*22*(1),1-12.

36. Afriyie, S. O., Addison, T. K., Gebre, Y., Mutala, A. H., Antwi, K. B., Abbas, D. A., ... &Badu,K.(2023).Accuracyof diagnosisamongclinicalmalariapatients:comparingmicroscopy, RDT and a highly sensitive quantitative PCR looking at the implications forsubmicroscopicinfections.*MalariaJournal*,*22*.

37. Imagegenerator(2023)mosquitoparasitehttps://gencraft.com/

38. Dukuzumuremyi,A.(2020). *MachinelearningbasedpredictionofmalariaoutbreakusingenvironmentdatainRwanda*(Doctoraldissertation,UniversityofRwanda).

39. Musoke, D., Atusingwize, E., Namata, C., Ndejjo, R., Wanyenze, R. K., & Kamya, M. R.(2023). Integrated malaria prevention in low-and middle-income countries: a systematicreview.*MalariaJournal*,*22*(1),79.

40. World Health Organization. (2023). *WHO guidelines for malaria, 14 March 2023* (No.WHO/UCN/GMP/2023.01).WorldHealthOrganization.

41. Mbunge,E.,&Batani,J.(2023).Applicationofdeeplearningandmachinelearningmodelsto improvehealthcareinsub-SaharanAfrica:Emergingopportunities,trendsandimplications.*TelematicsandInformaticsRe*

42. Symptoms of malaria (2023) https://www.who.int/news-room/fact-sheets/detail/malaria#:~:text=Symptoms%20can%20be%20mild%20or,%2C%20seizures%2C%20and%20difficulty%20breathing.

43. Minhaj,F.S.,Ogale,Y.P.,Whitehill,F.,Schultz,J.,Foote,M.,Davidson,W.,...&Wong, M. (2022).Monkeypoxoutbreak—Ninestates,May2022:Weekly/June10,2022/71(23);764–769.

44. Sayan, M., Sarigul Yildirim, F., Sanlidag, T., Uzun, B., Uzun Ozsahin, D., & Ozsahin, I.(2020).CapacityevaluationofdiagnostictestsforCOVID-19usingmulticriteriadecision-makingtechniques.*ComputationalandMathematicalMethodsinMedicine*,*2020*.

45. Mustapha, M. T., Ozsahin, D. U., Ozsahin, I., & Uzun, B. (2022). Breast cancer screeningbasedonsupervisedlearningandmulti-criteriadecision-making.*Diagnostics*,*12*(6),1326.

46. Dada, E. G., Oyewola, D. O., Joseph, S. B., Emebo, O., & Oluwagbemi, O. O. (2022).Ensemble machine learning for monkeypox transmission time series forecasting. *AppliedSciences*,*12*(23),12128.

47. UzunOzsahin, D.,Mustapha, M.T.,Uzun,B.,Duwa, B.,&Ozsahin,I.(2023).Computer-aided detection and classification of monkeypox and chickenpox lesion in human subjectsusingdeeplearningframework.*Diagnostics*,*13*(2),292.

48. Morang'a,C. M.,Amenga–Etego, L., Bah, S.Y.,Appiah, V., Amuzu, D.S.,Amoako, N., ...&Otto,T.D.(2020).Machinelearningapproachesclassifyclinicalmalariaoutcomesbasedonhaematological parameters.*BMC medicine*,*18*(1),1-16.

49. Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2022). Artificial intelligence in diseasediagnosis: a systematicliterature review, synthesizingframework and future researchagenda.*Journalofambientintelligenceandhumanizedcomputing*,1-28.

50. Brock,P.M.,Fornace,K.M.,Grigg,M. J.,Anstey, N.M.,William,T.,Cox,J.,...&Kao, R.R.(2019).Predictiveanalysisacrossspatialscaleslinkszoonoticmalariatodeforestation. *Proceedingsofthe RoyalSocietyB*,*286*(1894),20182351.

51. Mody, R. M., Murray, C. K., Dooley, D. P., Hospenthal, D. R., Horvath, L. L., Moran, K.A.,&Muntz,R.W.(2006).Theremotediagnosisofmalariausingtelemedicineore-mailedimages.*Militarymedicine*, *171*(12),1167-1171.

52. HierarchicalrepresentationofAI,MLandDL***https://builtin.com/machine-learning/what-is-deep-learning***

53. Hamsagayathri,P.,&Vigneshwaran,S.(2021,February).Symptomsbaseddiseaseprediction using machine learning techniques. In *2021 Third international conference onintelligent communication technologies and virtual mobile networks (ICICV)* (pp. 747-752).IEEE.

54. Katsaouni, N., Tashkandi, A., Wiese, L., & Schulz, M. H. (2021). Machine learning baseddiseasepredictionfromgenotypedata.*BiologicalChemistry*,*402*(8),871-885.

55. Terrada, O., Cherradi, B., Raihani, A., & Bouattane, O. (2020, April). Atherosclerosisdiseasepredictionusingsupervisedmachinelearningtechniques.In*20201stInternationalConference onInnovative Researchin Applied Science,Engineeringand Technology(IRASET)*(pp.1-5).IEEE.

56. Ali,M.M.,Paul, B.K.,Ahmed, K.,Bui, F.M.,Quinn,J.M.,&Moni, M.A.(2021).Heartdiseasepredictionusingsupervisedmachinelearningalgorithms:Performanceanalysisandcomparison.*Computers inBiologyandMedicine*, *136*,104672.

57. Aljanabi,M.,Qutqut,M.H.,& Hijjawi,M.(2018).Machinelearningclassificationtechniques for heart disease prediction: a review. *International Journal of Engineering &Technology*,*7*(4),5373-5379.

58. Gupta, C., Saha, A., Reddy, N. S., & Acharya, U. D. (2022). Cardiac Disease Predictionusing Supervised Machine Learning Techniques. In *Journal of Physics: Conference Series*(Vol.2161,No.1,p. 012013).IOPPublishing.

59. Dileep,P.,Rao,K.N.,Bodapati,P.,Gokuruboyina,S.,Peddi,R.,Grover,A.,&Sheetal, A. (2023). An automatic heart disease prediction using cluster-based bi-directional LSTM(C-BiLSTM)algorithm.*NeuralComputing andApplications*,*35*(10),7253-7266.

60. Antony,L.,Azam,S.,Ignatious,E.,Quadir,R.,Beeravolu,A.R.,Jonkman,M.,&DeBoer, F.(2021).Acomprehensiveunsupervisedframeworkforchronickidneydiseaseprediction. *IEEEAccess*,*9*,126481-126501.

61. Vankdothu, R., Hameed, M. A., & Fatima, H. (2022). A brain tumor identification andclassificationusingdeep learningbasedonCNN-LSTMmethod.*ComputersandElectricalEngineering*,*101*,107960.

62. Sharean, T. M., & Johncy, G. (2022). Deep learning models on Heart Disease Estimation-Areview.*JournalofArtificialIntelligence*, *4*(2),122-130.

63. Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ...&Kording,K.P.(2019).Adeeplearningframeworkforneuroscience.*Natureneuroscience*,*22* (11),1761-1770.

64. Ozsahin, D. U., Mustapha, M. T., Mubarak, A. S., Ameen, Z. S., & Uzun, B. (2022,August).ImpactofOutliersandDimensionalityReductiononthePerformanceofPredictive Models for Medical Disease Diagnosis. In *2022 International Conference onArtificialIntelligenceinEverything(AIE)*(pp.79-86). IEEE.

65. Motarwar, P., Duraphe, A., Suganya, G., & Premalatha, M. (2020, February). Cognitiveapproachforheartdiseasepredictionusingmachinelearning.In*2020internationalconferenceonemergingtrendsininformationtechnologyandengineering(ic-ETITE)*(pp.1-5).IEEE.

66. Onakpojeruo, E. P., Al-Turjman, F., Mustapha, M. T., Altrjman, C., & Ozsahin, D. U.(2022,October).Emerging AIandcloudcomputing paradigmsappliedtohealthcare.In *The International Conference on Forthcoming Networks and Sustainability (FoNeS2022)*(Vol.2022,pp.811-826). IET.

67. Ozsahin, D. U., Mustapha, M. T., Mubarak, A. S., Ameen, Z. S., & Uzun, B. (2022,August). Impactof feature scalingonmachinelearningmodelsforthe diagnosisofdiabetes. In *2022 International Conference on Artificial Intelligence in Everything (AIE)*(pp.87-94).IEEE.

68. Nelli,F.(2023). Machinelearning with scikit-learn.In *Python Data Analytics:WithPandas,NumPy,andMatplotlib*(pp.259-287).Berkeley,CA:Apress.

69. Wang, J., Rao, C., Goh, M., & Xiao, X. (2023). Risk assessment of coronary heart diseasebasedoncloud-randomforest. *ArtificialIntelligenceReview*,*56*(1),203-232.

70. Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., &Gonzales-Yanac,T.(2023).MultiplediseasepredictionusingMachinelearningalgorithms.*MaterialsToday:Proceedings*,*80*,3682-3685.

71. Stojanov,D.,Lazarova,E.,Veljkova,E.,Rubartelli,P.,&Giacomini,M.(2023).Predictingthe outcome of heart failure against chronic-ischemic heart disease in elderly population–Machine learning approach based on logistic regression, case to Villa ScassihospitalGenoa,Italy.*JournalofKingSaudUniversity-Science*,*35*(3),102573.

72. Sawhney, R., Malik, A., Sharma, S., & Narayan, V. (2023). A comparative assessment ofartificial intelligence models used for early prediction and evaluation of chronic kidneydisease.*DecisionAnalyticsJournal*,*6*,100169.

73. Khan, A., Qureshi, M., Daniyal, M., & Tawiah, K. (2023). A Novel Study on MachineLearning Algorithm-Based Cardiovascular Disease Prediction. *Health & Social Care in theCommunity*,*2023*.

74. Bhatt, C. M., Patel,P., Ghetia,T., & Mazzeo, P. L.(2023). Effectiveheartdiseasepredictionusingmachinelearningtechniques.*Algorithms*,*16*(2),88.

75. Priya, P. M. (2023, February). Disease Prediction and Classification using IntelligentAlgorithms. In *2023 Third International Conference on Artificial Intelligence and SmartEnergy (ICAIS)*(pp.494-498). IEEE.

76. Akash,S.,Abdelkrim,G.,Bayil,I.,Hosen,M.E.,Mukerjee,N.,Shater,A.F.,...&Tok, T. T. (2023). Antimalarial drug discovery against malaria parasites through haplopinemodification: An advanced computational approach. *Journal of Cellular and MolecularMedicine*.

77. Chen, J. X., & Zhao, X. Y. (2023). A general integrated machine learning pipeline: Itsconcept,main stepsandapplication inshearstrengthpredictionofRCbeamsstrengthenedwithFRCM.*EngineeringStructures*,*281*, 115749.

78. Singh, D., Halder, S., Bhattacharyya, S., Nath, I., Sahana, S., Pal, S., & Alkhafaji, M. A.(2023,September).COVID-19caseanalysis inIndiausingEDAanditsprediction.In*AIPConferenceProceedings*(Vol.2845,No.1).AIPPublishing.

79. Ibrahium, H. M., Saif Elnasr, M. K., & Shaker, H. A. (2023). Univariate analysis forprediction of factors affecting outcome of laparoscopic simple nephrectomy.*FayoumUniversity MedicalJournal*,*11*(3),149-167.

80. Sheng, N., Cui, H., Zhang, T., & Xuan, P. (2021). Attentional multi-level representationencodingbasedonconvolutionalandvarianceautoencodersforlncRNA–diseaseassociationprediction.*BriefingsinBioinformatics*,*22*(3),bbaa067.

81. Bizimana, P. C., Zhang, Z., Asim, M., El-Latif, A., & Ahmed, A. (2023). An EffectiveMachine Learning-Based Model for an Early Heart Disease Prediction. *BioMed ResearchInternational*,*2023*.

APPENDICES

AppendixA

EthicalApprovalDocument

Thereisnoethicalapprovaldocument thatcanbepresented.

Assoc.Prof.Dr.DilberUzunOzsahin
Supervisor

AppendixBCurriculum

Vitae

PersonalInformation

Surname,Name:BasilB.DuwaD
ate of Birth:01-10-
1990PlaceofBirth:Nigeria

TableB1.

Education.

Degree	Department/Program	University	YearofGraduation
B.Sc.	BiologicalSciences	Adamawa StateUniversity,Nigeria.	2018
M.Sc.	BiomedicalEngineering	NearEastUniversity	2021
PhD	BiomedicalEngineering	NearEastUniversity	Present

TableB2.

WorkExperience.

Title	Place	Year
Clinical research Assistant	Operationalresearchcenterinhealthcare,NEU	2021-present
Research Assistant	BiomedicalEngineeringDepartment	2020-2021
MilitaryTraining	National Youth Service to the Federal Republic ofNigeria.	2019
Monitoring andEvaluation(Data Analyst)	PlaninternationalMubiAdamawaState, Nigeria.	2018
Monitoring andEvaluation	SavetheChildrenInternational	2012

Publications in International Journals, Conferences and Book chapters

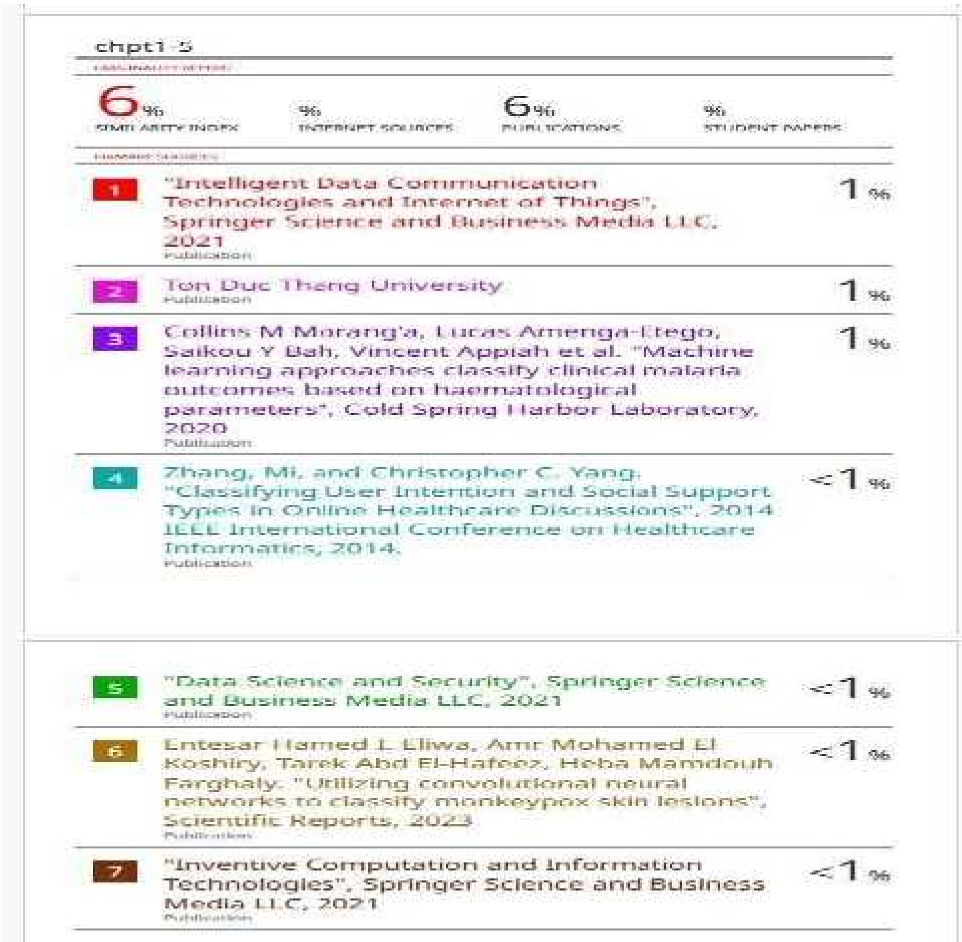
1. Uzun Ozsahin, D., Mustapha, M. T., Uzun, B., Duwa, B., & Ozsahin, I. (2023). Computer-aided detection and classification of monkeypox and chickenpox lesion in human subjects using deep learning framework. *Diagnostics*, *13*(2), 292.
2. Uzun Ozsahin, D., Mustapha, M. T., Bartholomew Duwa, B., & Ozsahin, I. (2022). Evaluating the performance of deep learning frameworks for malaria parasite detection using microscopic images of peripheral blood smears. *Diagnostics*, *12*(11), 2702.
3. Uzun Ozsahin, D., Precious Onakpojeruo, E., Bartholomew Duwa, B., Usman, A. G., Isah Abba, S., & Uzun, B. (2023). COVID-19 Prediction Using Black-Box Based Pearson Correlation Approach. *Diagnostics*, *13*(7), 1264.
4. Duwa, B. B., Kibarer, A., Uzun, B., Kaba, Ş., & Ozsahin, D. U. (2022, August). Evaluation of Techniques Used in Phenol Removal from Wastewater Using Fuzzy PROMETHEE Method. In *International Conference on Theory and Applications of Fuzzy Systems and Soft Computing* (pp. 289-296). Cham: Springer Nature Switzerland.
5. Ozsahin, D. U., Onakpojeruo, E. P., Duwa, B. B., Uzun, B., Ozsahin, I., & Chioma, E. C. (2023, February). Comparative Evaluation of 3D Filaments Used in Additive Manufacturing of Biomedical Tools; Using Fuzzy PROMETHEE. In *2023 Advances in Science and Engineering Technology International Conferences (ASET)* (pp. 1-7). IEEE.
6. Uzun, B., Uzun Ozsahin, D., & Duwa, B. (2021). Fuzzy Logic and Fuzzy Based Multi Criteria Decision Analysis. *Application of Multi-Criteria Decision Analysis in Environmental and Civil Engineering*, 47-56.
7. Ozsahin, D. U., Sheshakli, S., Kibarer, A. G., Denker, A., & Duwa, B. B. (2021). Analysis of early stage breast cancer treatment techniques. In *Applications of Multi-Criteria Decision-Making Theories in Healthcare and Biomedical Engineering* (pp. 71-80). Academic Press.
8. Ozsahin, D. U., Almoqayad, A. S., Ghader, A., Alkahlout, H., Idoko, J. B., Duwa, B. B., & Ozsahin, I. (2022). Development of smart jacket for disc. In *Modern Practical Healthcare Issues in Biomedical Instrumentation* (pp. 31-46). Academic Press.
9. Ozsahin, D. U., Idoko, J. B., Duwa, B. B., Zeidan, M., & Ozsahin, I. (2022). Construction of vehicle shutdown system to monitor driver's heartbeats. In *Modern Practical Healthcare Issues in Biomedical Instrumentation* (pp. 123-138). Academic Press.

&Ozsahin,I.(2022).Designinga3Dprintedartificialhand.In*ModernPractical*

HealthcareIssuesinBiomedicalInstrumentation(pp. 3-18).AcademicPress.

11. Ozsahin,D.U.,Emegano,D. I., Bader, B.A., Duwa,B. B., &Ozsahin,I.(2024). Bloodcircuitinhemodialysis.

AppendixCSimilarityReport



chpt1-5

ORIGINALITY REPORT

6%	%	6%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Intelligent Data Communication Technologies and Internet of Things", Springer Science and Business Media LLC, 2021 Publication	1%
2	Ton Duc Thang University Publication	1%
3	Collins M Morang'a, Lucas Amenga-Etego, Saikou Y Bah, Vincent Appiah et al. "Machine learning approaches classify clinical malaria outcomes based on haematological parameters", Cold Spring Harbor Laboratory, 2020 Publication	1%
4	Zhang, Mi, and Christopher C. Yang. "Classifying User Intention and Social Support Types in Online Healthcare Discussions", 2014 IEEE International Conference on Healthcare Informatics, 2014. Publication	<1%
5	"Data Science and Security", Springer Science and Business Media LLC, 2021 Publication	<1%
6	Entesar Hamed I. Eliwa, Amr Mohamed El Koshiry, Tarek Abd El-Hafeez, Heba Mamdouh Farghaly. "Utilizing convolutional neural networks to classify monkeypox skin lesions", Scientific Reports, 2023 Publication	<1%
7	"Inventive Computation and Information Technologies", Springer Science and Business Media LLC, 2021 Publication	<1%

Assoc.Prof.Dr.DilberUzunOzsahin

Supervisor