COVID-19 iVOICE iDIAGNOSIS

Sireesh ireddy1, iAbhinav iStephen2 i,Pranav iSteyn3,Devanjali iRelan4

*BML iMun ijal iUniversity, iGurgaon Haryana, iIndia*

*— i−*

E-mail: i[koduru.reddy.19cse@bmu.edu.in,](mailto:koduru.reddy.19cse@bmu.edu.in) imandala.stephen.19cse@bmu.edu.in i, imandala.steyn.19cse@bmu.edu.in, idevanjali.relan@bmu.edu.in

***Abstract i– i*Healthcare iprofessionals ihave iregularly iused iaudio isignals igenerated iby ithe ihuman ibody i(e.g., isighs, ibreathing, iheart, idigestion, ivibration isounds) ias iindicators ito idiagnose idisease ior iassess idisease iprogression. iUntil irecently, isuch isignals iwere iusually icollected imanually iduring ischeduled ivisits. iDigital itechnology iis inow ibeing iused iin iresearch ito icollect ibodily isounds i(e.g., ifrom idigital istethoscopes) ifor icardiovascular ior irespiratory iexamination, iwhich ican ithen ibe iused ifor iautomatic ianalysis. iSome ipreliminary iresearch ishows ipromise iin idetecting iCOVID-19 idiagnostic isignals iin ivoice iand icoughs. iWe idescribe iour idata ianalysis iover ia ilarge-scale icrowdsourced idataset iof irespiratory isounds icollected ito iaid iin ithe idiagnosis iof iCOVID-19 iin ithis ipaper. iCoughs iand ibreathing iare iused ito idetermine ihow idistinguishable iCOVID-19 isounds iare ifrom ithose iin iasthma ior ihealthy icontrols. iOur ifindings ishow ithat ieven ia isimple ibinary imachine ilearning iclassifier ican icorrectly iclassify ihealthy iand iCOVID-19 isounds.**

***Keywords i– i*COVID-19, iCrowdsourcing iPlatform, iAudio iAnalysis, iCoughing, iBreathing.**

# INTRODUCTION

Clinicians iand iclinical iresearchers ihave ifrequently iused iaudio isignals igenerated iby ithe ihuman ibody i(e.g., isighs, ibreathing, iheart, idigestion, ivibration isounds) iin idisease idiagnosis iand imonitoring. iUntil irecently, ihowever, isuch isignals iwere itypically icollected ithrough imanual iauscultation iduring ischeduled ivisits. iDigital itechnology iis inow ibeing iused iin iresearch ito icollect ibodily isounds i(e.g., idigital istethoscopes) iand irun iautomatic ianalysis ion ithe idata i, ifor iexample, ifor iwheeze idetection iin iasthma i. iResearchers ihave ialso ibeen itesting ithe iuse iof ihuman ivoice ito iaid iin ithe iearly idiagnosis iof ia ivariety iof iillnesses: iParkinson’s idisease iis iassociated iwith isoftness iof ispeech i(due ito ia ilack iof icoordination iof ithe ivocal imuscles), ivoice ifrequency iwith icoronary iartery idisease i(hardening iof ithe iarteries, iwhich imay iaffect ivoice iproduction) i, iand ivocal itone, ipitch, irhythm, irate, iand ivolume iwith iinvisible iillnesses isuch ias ipost-traumatic istress idisorder i. iThe iuse iof ihuman-generated iaudio ias ia ibiomarker ifor ivarious iillnesses iholds ienormous ipromise ifor iearly idiagnosis ias iwell ias iaffordable isolutions ithat icould ibe imade iavailable ito ithe igeneral ipublic iif iembedded iin icommodity idevices. iThis iis iespecially itrue iif isuch isolutions ican imonitor ipeople iin itheir idaily ilives iin ian iauthentic imanner. iRecent iresearch ihas ibegun ito iinvestigate ihow irespiratory isounds i(e.g., icoughs, ibreathing, iand ivoice) icollected iby idevices ifrom iCOVID-19 ipositive ihospital ipatients idiffer ifrom isounds ifrom ihealthy ipeople. iIn

, idigital istethoscope idata ifrom ilung iauscultation iis iused ias ia idiagnostic isignal ifor iCOVID-19; iin, ia istudy iof iCOVID- i19 icough idetection iusing iphone idata iis ipresented iusing ia icohort iof i48 iCOVID-19 ipatients iversus iother ipathological icoughs ion iwhich ian iensemble iof imodels iis itrained. iSpeech irecordings ifrom ihospital ipatients iwith iCOVID-19 iare ianalyzed iin ito iautomatically icategorize ipatients’ ihealth istates. iIn iour iwork, iwe iinvestigate ithe iuse iof ihuman irespiratory isounds ias idiagnostic imarkers ifor iCOVID-19 iin icrowdsourced, iuncontrolled idata. iThis ipaper,in iparticular, idescribes iour ipreliminary ifindings ifrom ia isubset iof iour idataset, iwhich iis icurrently ibeing icrowdsourced iglobally iat iwww.covid-19- isounds.org. iThe idata iwas igathered iusing ian iapp i(Android iand iWeb) ithat iasked ivolunteers ito iprovide isamples iof itheir ivoice, icoughs, iand ibreathing, ias iwell ias itheir imedical ihistory iand

symptoms.

The iapp ialso iinquires ias ito iwhether ithe iuser ihas itested ipositive ifor iCOVID-19. iTo idate, iwe ihave icollected iapproximately i10,000 isamples ifrom iapproximately i7000 iunique iusers. iThis iis, ito iour iknowledge, ithe ilargest iuncontrolled, icrowdsourced idata icollection iof iCOVID-19 irelated isounds iworldwide iWe ipresent ipreliminary ifindings ifor iCOVID-19 ion ithe idiscriminatory ipower iof icoughs iand ibreath isounds. iWe idevelop ithree ibinary iactivities: ione ito idifferentiate iCOVID-19 ipositive iusers ifrom ihealthy iusers; ione ito idifferentiate iCOVID-19 ipositive iusers iwith ia icough ifrom ihealthy iusers iwith ia icough; iand ione ito idifferentiate iCOVID-19 ipositive iusers iwith ia icough ifrom iusers iwith iasthma iwho ireport ihaving ia icough. iThe iresults ishow ithat iperformance ifor iall iactivities iremains iabove i80 i% iArea iUnder iCurve i(AUC). iWe iare iable ito icorrectly iclassify ihealthy iand iCOVID-19 isounds iwith ian iAUC iof i80%. i(Activity i1). iWhen iattempting ito idistinguish ia iuser iwho itested ipositive ifor iCOVID-19 iand ihas ia icough ifrom ia ihealthy iuser iwith ia icough i((Activity i2), iour iclassifier iachieves ian iAUC iof i82%, iwhereas iwhen iattempting ito idistinguish iusers iwho itested ipositive ifor iCOVID-19 iand ihave ia icough ifrom iusers iwith iasthma iand ia icough i((Activity i3), iwe iachieve ian iAUC iof i80%.

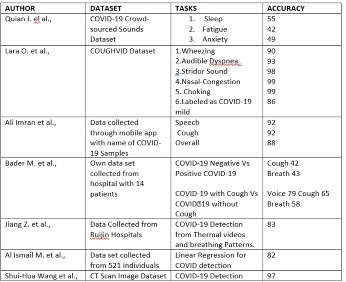
* We idemonstrate ihow iwe ican iuse iaudio idata iaugmentation ito iimprove ithe irecall iperformance iof isome iof iour iactivities iwith iless idata. iWe isee ia i5% iand i8% iimprovement iin iperformance ifor i(Activities i2 iand i3, irespectively.
* Discussion iof iresults iand itheir ipotential, ias iwell ias iillustration iof iseveral ifuture idirections ifor iour ianalysis iand isound-based idiagnostics iin ithe icontext iof iCOVID-19, iwhich icould ilead ito iCOVID-19 ipre-screening iand iprogression idetection.

# 1 iMOTIVATION iAND iRELATED iWORK

Over i521 imillion icases iof ithe inovel icoronavirus idisease ihave ibeen ireported isince iits ioutbreak, iwith i62 imillion ideaths. iResearchers iand iscientists ihave imade isignificant iprogress iin ideveloping iCOVID-19 itreatments iand ivaccines, iand ieffective iand ieasily iaccessible itests ihave ibeen icritical iin iquickly itracing iinfected ipeople. iThe ireverse itranscription ipolymerase ichain ireaction i(RT-PCR) iassay ito idetect ithe ipresence iof iviral iribonucleic iacid i(RNA) ifrom iswab isamples iis icurrently ithe imost icommonly iused iand ifirst-line idiagnostic itool ifor iCOVID-

19. iRT-PCR itests iare ihighly isensitive iin ithe ilaboratory i(over i95 i% idiagnostic isensitivity iand ispecificity),but ihave ibeen ifound ito iperform idifferently iin icommercial ikits, iwith isensitivity iranging ifrom i75 ito i100 ipercent, iand iin ithe iworst icase ireaching ias ilow ias i38 i%. iFurthermore, ithe isample ianalysis iprocedure iis iinvolved. itime iconsuming, iand ionly iavailable iin iapproved ilaboratories ihighly itrained ipersonnel, iresulting iin ilimited itesting icapacity iand ifailure ito ikeep iup iwith ithe irapid iincrease iin idemand icomputer iTomography i(CT). iScanners iare ibecoming imore ipopular ifor iCOVID-19 idiagnostics iin isome iareas ilike iChina.

However, iThis imethod, ihad inot ibeen itried ibecause ithere iare istill ia ilot iof idoctors iaround iis iskeptical iof ithe ireported ihigh isensitivity.

In iaddition, iCT iscanners iare ispecialized iand iexpensive iequipment isuitable ionly ifor imedical icenters iwith itrained istaff ifor iits ioperation. i i iFor iinpatients, ion itop iof ia ihigh iprice itag ifor ia isingle iscan, ipatient itransport ito iand ifrom ithe iscanner irequires ito ibreak ithe iisolation, iwhich isignificantly iincreases ithe iinfection itransmission irisk. iIt iis icritical ithat ithe ipandemic iresponse iovercomes ithe ilimitations iof iRT-PCR iand iCT ito itest ion ia ilarge iscale iin ia itimely imanner. iThis irequires iquick, ilow-cost, ilong-term, iand ieffective itesting imethods ithat ican ibe irepeat iover itime ito itrack iprogress. iThis iwould inot ionly ihelp ito icontain ithe icurrent ispread, ibut iit iwould ialso isuppress iresurgence iand ireduce ihealth irisks.

Machine ilearning imethods ihave ibeen ideveloped ito irecognize iand idiagnose irespiratory idiseases ibased ion isounds, iparticularly icoughs iutilizes iconvolutional ineural inetworks i(CNN’s) ito idetect icough iin iambient iaudio iand idiagnose ithree ipotential iillnesses i(bronchitis, ibronchiolitis, iand ipertussis) ibased ion itheir idistinct iaudio ifeatures. iSeveral imodels ifor iCOVID-19 iprediction iusing iaudio ihave ibeen ideveloped iand ipublished iin ithe ilast iyear iin ithis icontext. iMachine ilearning iadvances ihave idemonstrated ithe ipotential iof iautomated iauscultation iof irespiratory isounds iand iopened iup inew iopportunities ifor ifully iautomated iCOVID-19 iscreening.

# METHODOLOGY

Feature-based imachine ilearning iand ishallow iclassifiers iwere iused idue ito ithe imoderate isize iof ithe idataset ichosen iand ithe irelevance iof iexplainability igiven ithe ipublic ihealth iimplications iof iour iresearch. iwe iadopted istandard idata iprocessing iand imodeling ipractices ifrom ithe iaudio iand isound iprocessing iliterature ifor imedical ipurposes. iWe idetail ithe iextracted ifeatures iand ithe iapproach iwe iused ito itrain irobust iclassification imodels iin ithis ipart, itaking iinto iconsideration ithe iunique icharacteristics iof iour idata i(e.g., ilongitudinal imobile iusers iand icross-validation). iWe ilooked iat icharacteristics ithat iwere icreated iand ithose ithat iwere ilearned ithrough itransfer ilearning. iWe iput iLogistic iRegression i(LR), iGradient iBoosting iTrees i(GBTs), iand iSupport iVector iMachines i(SVMs) ito ithetest; ithe iresults imay ibe ifound iin ithe iresults isection. iAn iSVM iclassifier iusing ia iRadial iBasis iFunction i(RBF) ikernel iwas itested.

1. *Dataset iused ifor ithis ianalysis*

We ihave iobtained ithe idataset ifrom ithe iUniversity iof iCambridge iusing ione-to-one ilegal iagreements ifor iresearch ipurposes, idue ito ithe isensitive inature i(e.g. ivoice) iof ithe idataset.The ititle iof ithe ipaper ishould ibe ias isuccinct ias ipossible, istating ithe isubject iof ithe ipaper iin ia ivery iclear imanner. iIt ishould ibe icentered iat ithe itop iof ithe ifirst ipage, iin ibold, itype isize i14 ipoints, iwith ithe iwhole ititle iin icapital iletters.

This idataset ifocused ion ia icurated icollection iof ithe iacquired idata ifor ithis iresearch, iguided imostly iby ithe iimbalance iof iCOVID-19 itested iparticipants iin ithe idataset i(until i22 iMay i2020). iThey ialso ilimited iour iwork ito icoughs iand ibreathing isolely i(and inot ithe ivoice isamples). iOne iaudio irecording iis irepresented iby ia isample. iAfter ifiltering i(silent iand inoisy isamples), iwe ipresent ithe inumber iof isamples iused iin iour istudy. iThey iextracted iand icarefully iverified iall isamples ifrom iindividuals iwho iindicated ithey ihad itested ipositive ifor iCOVID- i19 i(during ithe ilast i14 idays ior ibefore), itotaling i141 icough iand ibreathing isamples. iThere iwere i54 isamples ifrom iusers iwho ihad ia idry ior iwet icough.

Three isets iof iusers iare iused ias ia icontrol igroup iin iour istudy. iUsers ifrom icountries iwhere ithe ivirus iwas inot iprevalent iat ithe itime iof idata icollection i(up ito iaround i2000 icases) iare iclassified ias inon-COVID iusers. iAlbania, iBulgaria, iCyprus, iGreece, iJordan, iLebanon, iSri iLanka, iTunisia, iand iVietnam iwere iamong ithe icountries iwe ichose. iNon-COVID iusers iare idefined ias ipeople iwho ihave ia iclean imedical ihistory, ihave inever ismoked, ihave inot itested ipositive ifor iCOVID19, iand ihave inot ireported iany isymptoms. i298 isamples iwere icontributed iby ithese iusers. iThe inon-COVID iwith icough igroup iconsisted iof iindividuals iwho imet ithe isame icriteria ias ithe inon-COVID igroup ibut ireported icough ias ia isymptom; ithis igroup iprovided i32 isamples. iFinally, iasthma iwith icough iusers ihad iasthma, ihad inot itested ipositive ifor iCOVID-19, iand icoughed; ithey iprovided ius iwith i20 isamples.

1. *Pre-Processing*
   1. *Feature iextraction:*

*i***Handcrafted iFeatures: i**The iapps’ iraw isound iwave iforms iare ire-sampled ito i22kHz, iwhich iis ia icommon ivalue ifor iaudio ioperations. iOur iaudio iprocessing ilibrary iwas ilibrosa i. iAt ithe iframe iand isegment ilevel, inumerous ihandmade icharacteristics iincluding ifrequency- ibased, istructural, istatistical, iand itemporal iproperties iare iretrieved ifrom ithe ire-sampled iaudio. iA isegment iis ithe ientirety iof ia isingle iaudio irecording, iwhereas ia iframe iis ia iportion i(subset) iof ithe ientire iaudio idata iincluded iin ia isegment.

* **Duration**: ithe itotal iduration iof ithe irecording iafter itrimming ileading iand itrailing isilence.
* **Onset**: ithe inumber iof ipitch ionsets i(pseudo isyllables) iis icomputed ifrom ithe isignals, iby iidentifying ipeaks ifrom ian ionset istrength ienvelope, iwhich iis iobtained iby isumming ieach ipositive ifirst-order idifference iacross ieach iMel ibandoccur iat iregular itemporal iintervals. iIn iour icontext, iit iis iused ifor iits ipeak idetection icapabilities.
* **Spectral iCentroid**: ithe imean i(centroid) iextracted iper iframe iof ithe imagnitude ispectrogram.
* **Roll-off iFrequency**: ithe icenter ifrequency ifor ia ispectrogram ibin iso ithat iat ileast i85% iof ithe ienergy iof ithe ispectrum iin ithis iframe iis icontained iin ithis ibin iand ithe ibins ibelow.
* **MFCC**: iMel-Frequency ifrom ithe ishort-term ipower ispectrum, ibased ion ia ilinear icosine itransform iof ithe ilog ipower ispectrum ion ia inonlinear iMel iscale. iMFCCs iare iamongst ithe imost icommon ifeatures iin iaudio iprocessing. iWe iuse ithe ifirst i13 icomponents.
* **Period**: ithe imain ifrequency iof ithe ienvelope iof ithe isignal. iWe icalculate ithe iFFT ion ithe ienvelope iand iidentify ithe ifrequency iwith ithe ihighest iamplitude ifrom ithe i4th imode iupwards i(as ithe ienvelope ihas inon-zero imean).
* **RMS iEnergy**: ithe iroot-mean-square iof ithe imagnitude iof ia ishort-time iFourier itransform iwhich iprovides ithe ipower iof ithe isignal.
* Δ**-MFCC**: ithe itemporal idifferential i(delta) iof ithe iMFCC.

.

We iextract imany istatistical icharacteristics ifor ithe ifeatures ithat igenerate itime iseries i(RMS iEnergy, iSpectral iCentroid, iRoll-off iFrequency, iand iall iversions iof iMFCCs) ito icapture ithe idistributions ibeyond ithe imean. iMean, imedian, iroot- imean-square, imaximum, iminimum, i1st iand i3rd iquartile, iinterquartile irange, istandard ideviation, iskewness, iand ikurtosis iare iall iincluded iin ithe ilist. i i i i iThe ifirst ifour isegment- ilevel ifeatures, ifour iframe-level ifeatures irepresented iby itheir istatistics, iand ithree ivariations iof iMFCCs iwith ieach icomponent irepresented iby iits istatistics itotal i477 ihandcrafted ifeatures i(4 i+ i4 i× i11 i+ i3 i× i13 i× i11 i= i477).

1. *Algorithm i/Model iArchitecture*

**Features ifrom iTransfer iLearning**. iWe iuse iVGGish ito iextract iaudio ifeatures iautomatically iin iaddition ito ihandcrafted ifeatures. iVGGish iis ia itype iof iconvolutional ineural inetwork. iAudio iclassification ibased ion iraw iaudio iinput iwas iproposed. iA ilarge-scale iYouTube idataset iwas iused ito itrain ithe iVGGish imodel. iThe imodel iparameters ithat iwere idiscovered iwere imade iavailable. iIt’s isomething iwe iuse ias ia ifeature iextractor ifor iconverting iraw iaudio iwave iforms ito iembeddings i(features) ithat iare isubsequently iused ito itrain ia ishallow ineural inetwork iclassifier. iThe iVGGish ipre-trained imodel iseparates idata isamples iinto i0.96-sec inon-overlapping isub-samples iand ioutputs ia i128-dimensional ifeature ivector ifor ievery i0.96 iseconds. i16 ikHz iis ithe isampling irate. iThe ifinal ifeatures iare ithe imean iand istandard ideviation iacross ithe ientire isegment, iwith idimension i256 i(128x2). iBecause iVGGish ionly iaccepts ia ispectrogram ias iinput, isome ikey itemporal iproperties imay ibe iignored iin ithe ifeature ispace, inecessitating ithe iemployment iof ia icombination iof iVGGish iand ihandcrafted ifeatures. iSection i5 idemonstrates ithat ithis icombination iimproves iAUC iwhen icompared ito iutilizing ieither iVGGish ior ihandmade ifeatures.

We iget ia i477-dimensional ihandcrafted ifeature ivector, ia i256-dimensional iVGGish-based ifeature ivector, iand imany icomposite ifeature ivectors itotaling i733 idimensions ifor ieach imodality i(cough, ibreathing). iEach icombined ifeature ivector iis imade iup iof ia isubset iof ithe ihandmade ifeature isets iand iVGGish- ibased ifeatures iconcatenated itogether. iPrincipal iComponents iAnalysis i(PCA) ifurther ireduces ithese ifeature ivectors iwhile imaintaining ia iportion iof ithe iinitially iexplained ivariance.

# EVALUATION

The iclassification iof iaudio isamples ias iCOVID-19 ior ihealthy iusing ithe ifeatures igiven iin iSection i4 iis inow idetailed. iA isubset iof ithe ioriginally igathered idataset i(detailed iin iSection i3.3) iwas iused idue ito ithe ihigh-class iimbalance. iWe ibegin iby iexplaining ihow idata ifrom ivarious imodalities iwas icombined iand ithe idataset iwas ipartitioned ifor ithe iexperiments. iThe iexperiments, ias iwell ias ithe iAndroid iand iiOS iapp icodebases, iare ifreely iavailable ito iencourage irepeatability. iThe isection iconcludes iwith ia idiscussion iof ithe ifindings iand ioutcomes.

1. *Experimental isetup*
   1. *Classification iactivities. i: i*We ifocus ion ithree iclinically iimportant ibinary iclassification iactivities ibased ion ithe idata icollected i(Section iIII-A) i

**Activity i1**: iSeparate iusers iwho ihave ideclared ithey itested ipositive ifor iCOVID-19 i(COVID-positive) ifrom iusers iwho ihave inot ideclared ia ipositive itest ifor iCOVID- i19, ihave ia iclean imedical ihistory, ihave inever ismoked, ihave ino isymptoms, iand iwere iin icountries iwhere iCOVID-19 iwas inot iprevalent iat ithe itime, ias idescribed iin iSection i3. i(non-COVID). iWhile iwe icannot iguarantee ithat ithey iwere inot icontaminated, ithe ichances iare islim. i

**Activity i2**: iSeparate iusers iwho ihave ideclared ithey itested ipositive ifor iCOVID-19 iand ihave ilisted icough ias ia isymptom i(a icommon isymptom iin ipersons iwith iCOVID, ias ishown iin iFigure i3), i(COVID-positive icough), ifrom iusers iwho ihave iclaimed ithat ithey ihave inot itested ipositive ifor iCOVID-19, ithat ithey ihave ia iclean imedical i

background, ithat ithey ihave inever ismoked, ithat ithey ilived iin iplaces iwhere iCOVID-19 iwas inot iprevalent iat ithe itime, iand ithat ithey ihave ia icough ias ia isymptom i(non-COVID iwith icough). i

**Activity i3**: iSeparate iusers iwho ihave ideclared ithey itested ipositive ifor iCOVID-19 iand ihave icough ias ia isymptom i(COVID-positive iwith icough) ifrom iusers iwho ihave inot ideclared ithey itested ipositive ifor iCOVID-19, iare ifrom icountries iwhere iCOVID-19 iwas inot iprevalent iat ithe itime, ihave iasthma iin itheir imedical ihistory, iand ihave icough ias ia isymptom i(COVID-positive iwith icough) i(non-COVID iwith icough).

* 1. *Data iexploration: i*We ievaluate ithe idifferences ibetween ithe idistributions iof ithe ifeatures iderived ifrom icough iand ibreathing isubdivided iby irespective iclass ias ia ifirst istep ifollowing ifeature iextraction. iWe icannot iprovide iall idistributions idue ito ithe ilarge idimensionality iof ithe ifeatures, ithus iwe ifocus ionly ion ithe imean istatistical icharacteristic iof ieach ifeature ifamily i(e.g., iCentroid iis iCentroid imean ihere). iCoughs iand ibreaths ifrom iCOVID-positive iusers ihad ilonger itotal idurations, imore ionsets, ihigher iperiods, iand ilower iRMS, ibut itheir iMFCC ifeatures i[first icomponent iand ideltas] ihave ifewer ioutliers, ias ishown iin iFigure i5. iCOVID-positive iusers’ isamples iconcentrate imore itowards ithe imean iof ithe idistributions iin iboth iactivities, iwhereas ithe igeneral i(healthy) ipopulation idisplays ia iwider irange i(interquartile irange), iThe itheory iis ithat ia i(perhaps iforced) ihealthy icough iand ibreathing iare iextremely ivaried. iThis icould ialso iimply ithat icoughs iand ibreaths iare ieffective isounds ifor iidentifying iCOVID iand inon- iCOVID iusers.
  2. *Feature iablation istudies: i*We irepeat iour istudies iwith ithree idifferent iaudio iinputs: ionly icough, ionly ibreathing, iand imixed ito isee iwhich iaudio imodality i(cough ior ibreathing) icontributes imore ito icategorization iperformance. iWe irun itests ito idiscover ithe iappropriate icut-off ivalue ifor iPCA ito iaccount ifor ithe irising idimensionality iof ithe icomposite irepresentation iand ito imake ia ifair icomparison i(see iresults iin inext isection). iThe ipercentages iof iexplained ivariance iare i[70%, i80%, i90%, iand i95 i%]. iThis imeans ithat ithe iclassifiers iwill irequire ifewer iInput idimensions iif ithe iexplained ivariance iis ilower, iand ivice iversa. iTo iavoid iover ifitting, ia imixed irepresentation imay irequire ia imore icompressed irepresentation ithan ia irepresentation ibased isolely ion icoughs ior ibreaths.
  3. *Cross-validation iwith iusers: i*We imake itraining iand itest isets ifrom idisjoint iuser idivides, iensuring ithat ino isamples ifrom ithe isame iuser iappear iin iboth isplits. iIt iis iimportant ito inote ithat iapproach idoes inot iproduce iprecisely ibalanced iclass isplits; ihowever, iwhen inecessary, iwe idown isampled ithe imajority i(non-COVID) iclass. iThe ibalance iof ithe itest iset iis imaintained.

Even iso, iit’s idifficult ito iensure ithat ia isplit iselects ia irepresentative itest-set, itherefore iwe iused ia i10-fold-like icross ivalidation iin ithe iouter iloop i(80% i/20% isplit) iand ia ihyper-parameter isearch iin ithe iinner iloop ito idiscover ithe ibest iparameters i(using ithe i80% itrainset iin ia i5-fold icross ivalidation). iThis iconfiguration iis isimilar ito inested icross- ivalidation. iWe itest i5400 imodels i(3 iActivities i× i3 imodalities i× i10 iuser isplits i× i4 idimensionality ireduction icut-offs i× i3 ifeature itypes i× i5 ihyper-parameter icross- ivalidation iruns). iWe iused ithe iReceiver iOperating iCharacteristic i- iArea iUnder iCurve i(ROC- iAUC), iPrecision, iand iRecall ias istandard ievaluation icriteria. iThe istandard ideviation iand iaverage iperformance iof ithe iouter ifolds i(10 iuser-splits) iare ireported. iThe iperformance iof iour ithree iactivities iis ireported iin ithe inext isection.

* 1. *Sensitivity ito idemographics: i*Including iage iand isex ias ione-hot-encoded ifeatures iin iour imodels i(for iexample, iage igroup: i40-49 iyears iold) ihad ino isignificant ieffect ion ithe iresults i(< i± i2 iAUC).

1. *Distinguishing iCOVID-19 iusers ifrom ihealthy iusers*

Table i1 ishows ithe iclassification iresults ifor ithe ithree iactivities imentioned iearlier. iWe ipresent ithe ibest iresults ifor ieach iactivity, iwhich imay ihave ibeen iacquired iusing ia isingle imodality i(cough ior ibreathing inoises) ior ia icombination iof iboth. iThe ifirst irow ishows ithe iclassification iresults ifor iActivity i1: ithe ibinary

Table

Description automatically generated

iclassification iactivity iof idistinguishing iusers iwho istate ithat ithey ihave itested ipositive ifor iCOVID-19 i(COVID- ipositive) ifrom ithose iwho idid inot i(non-COVID). iAccording ito ithe imetrics, ithere iappear ito ibe isome idiscriminatory isignals iin ithe idata, iimplying ithat iuser icoughs ipaired iwith ibreathing icould ibe ia igood ipredictor iwhen iscreening ifor iCOVID-19. iThe iAUC ifor ithis ijob iis iat i80%, iwhile iprecision iand irecall iare iaround i70%. iActivity i1 ihas ithe ilowest istandard ideviations iacross ithe iuser-splits iwhen icompared ito ithe iother iactivities i(Activity i2 iand i3), iowing ito ithe ihigher idata iset. iWe iused ia isimple iclassifier i(Logistic iRegression), iand ithe idata iwas iperhaps itoo ismall ito ieliminate ithe inoise iand ivariability ibrought iby iour icrowd isourced idata icollection i(e.g., idifferences iin imicrophones, isurrounding inoises, iways iof iinputting ithe isounds). iNonetheless, ithese ifindings iprovide ius iwith iconfidence iin ithe isignal’s istrength. iWe ialso idiscovered ithat iwhen ihandcrafted ifeatures iare icombined iwith ifeatures ilearned ivia iVGGish, ithe iresults iare ibetter ithan iwhen ihandcrafted ior itransfer ilearning ifeatures iare iused ialone, idemonstrating ithe ivalue iof iapplying itransfer ilearning iin iour iresearch.

1. *Distinguishing iCOVID-19 icoughs ifrom iother icoughs*

The ibinary iclassification iof iusers iwho ireported itesting ipositive ifor iCOVID-19 iand ialso ideclared ia icough, ias iwell ias ia isimilar inumber iof iusers iwho iindicated ithey idid inot itest ipositive ifor iCOVID-19 ibut ideclared ia icough, iis idescribed iin ithe isecond irow iof iTable i1. i(Activity i2). iAn iAUC iof i82 ipercent iis ithe ibest iresult. iThis ijob ihas ian iaccuracy iof i80%, iindicating ithat icough inoises ican idistinguish iCOVID-19 ipositive iusers irather iwell. iThis imodel ihas ia ilittle ilower irecall i(72 ipercent), iindicating ithat iit icasts ia igood ibut isomewhat ispecialized inet:

it idoes inot icatch ievery iCOVID-19 icough, ibut iit idoes idetect ia ilot iof ithem. iNonetheless, igiven ithe iamount iof ithe idata iand ithe irelatively ihigh istandard ideviations icompared ito iActivity i1, irenders ithis iresult ipreliminary. iUsers iwho istated ithey idid inot itest ipositive ifor iCOVID-19 ibut ihad iasthma iand ideclared ia icough iwere icompared ito iusers iwho isaid ithey idid inot itest ipositive ifor iCOVID-19 ibut ihad iasthma iand ideclared ia icough, ias iindicated iabove. iTable i1’s ilast irow iindicates ian iAUC iof i80%. iWhile

irecall iis iacceptable, iprecision iis ilikewise ihigh ifor ithis iassignment, ias iit iis ifor ithe iother itwo. iBreathing inoises iappear ito ibe imore ipotent isignals ifor idiscriminating iusers iin ithis iactivity, iwhich iis iintriguing. iWe iinvestigated ithe iuse iof idata iaugmentation ito iincrease iperformance ion iActivities i2 iand i3.

# DISCUSSION iAND iCONCLUSIONS

We idiscussed ian iongoing iinitiative ito icrowd-source irespiratory isounds iand iinvestigate ihow isuch idata icould ihelp idiagnose iCOVID-19. iThese ifindings imerely iscrape ithe isurface iof ithis itype iof idata’s ipotential; iwhile ithey iare iencouraging, ithey iare inot istrong ienough ito ibe iused ias ia isingle iscreening itool. iTo ideal iwith ithe iissue ithat ithe ifraction iof iCOVID-19 ipositive iusers iis imodest, iwe’ve iconfined iourselves ito iusing ia isubset iof ithe idata iacquired ifor ithe itime ibeing. iWe ialso ihad ino iground itruth ion ihealth istatus, ithus iwe iassumed iusers ifrom icountries iwhere iCOVID-19 iwas inot iwidespread iat ithe itime iwere ihealthy iwhen iself-reporting ias isuch i(however, ithis ilimited iour idataset ifurther).

We imentioned ia icurrent iproject ito icrowd isource irespiratory isounds iand ilook iinto ihow ithis iinformation icould ihelp iidentify iCOVID-19. iThese ifindings ionly iscratch ithe isurface iof ithe ipotential iof ithis itype iof idata; iwhile iencouraging, ithey iare iinsufficient ito ibe iemployed ias ia isingle iscreening itool. iTo ideal iwith ithe iissue iof ia ismall ipercentage iof iCOVID-19 ipositive iusers, iwe’ve ilimited iourselves ito iusing ia isubset iof ithe idata ifor ithe itime ibeing. iWe ialso ihad ino iground itruth ion ihealth istatus, itherefore iwe iassumed iusers ifrom icountries iwhere iCOVID-19 iwas inot iwidely iused iat ithe itime iwere ihealthy iwhen iself-reporting i(however, ithis ilimited iour idataset ifurther). iWhile iwe ionly ilooked iat ithe idifference ibetween icough isounds iin iCOVID-19 iand iasthma, iour idataset iincludes ipeople iwith idifferent irespiratory idiseases, iand iwe iplan ito ilook iinto ithis ifurther ito isee ihow idistinguishable iCOVID-19 iis iin ithis iregard. iBecause ithe imobile iapp iencourages iusers ito iproduce isamples ievery icouple iof idays, iwe ihave imany iusers ifor iwhom iwe ican iexamine ithe ievolution iof irespiratory isounds iover itime. iThis iis iextremely iimportant ifor iCOVID-19, iand iit’s isomething iwe ihaven’t iinvestigated iyet iin iour ipresent iresearch.

Finally, iour imobile iapp ijust igathers idata iand idoes inot iprovide imedical iadvice; iwhile iwe ibelieve ithat ithe imodels icreated ifrom ithis idata imay ibe ibeneficial iin iillness iscreening, iwe iare iaware iof ithe idifficulties iinvolved iin iproviding imedical iadvice ito iusers iand ithe iarguments ithat ithis igenerally icauses.

# ACKNOWLEDGEMENTS

We ithank iDr.Devanjali iRelan ifor iencouraging ito ichoose ithis itopic iright ifrom ithe ibeginning. iHer iteachings iand iadvises iwere ivery ihelpful ito ius ito icomplete iour iproject.

# REFERENCES

1. 2019. ilibrosa.feature.delta ihttps://librosa.github.io/librosa/generated/librosa. ifeature.delta.html, inote= iAccessed: i2020-05-30.
2. 2020. icoughvid.

https://coughvid.epfl.ch/about/, inote= iAccessed: i2020- i05-30

1. 2020. iDetect iNow

https://detectnow.org/, inote= iAccessed: i2020-05-30

1. Charles iBales, iMuhammad iNabeel, iCharles iN. iJohn, iUsama iMasood, iHaneya iN. iQureshi, iHasan iFarooq, iIryna iPosokhova, iand iAli iImran. i2020. iCan imachine ilearning ibe iused ito irecognize iand idiagnose icoughs? iarXiv:2004.01495 i[eess.AS] i10 ipages.
2. Debrup iBanerjee, iKazi iIslam, iKeyi iXue, iGang iMei, iLemin iXiao, iGuangfan iZhang, iRoger iXu, iCai iLei, iShuiwang iJi, iand iJiang iLi. i2019. iA ideep itransfer ilearning iapproach ifor iimproved ipost-traumatic istress idisorder idiagnosis. iKnowledge iand iInformation iSystems i60, i3 i(2019), i1693–1724.
3. L iBrabenec, iJ iMekyska, iZ iGalaz, iand iIrena iRektorova. i2017. iSpeech idisorders iin iParkinson’s idisease: iEarly idiagnostics iand ieffects iof imedication iand ibrain istimulation. iJournal iof iNeural iTransmission i124, i3 i(2017), i303–334.
4. Gavin iC iCawley iand iNicola iLC iTalbot. i2010. iOn iover-fitting iin imodel iselection iand isubsequent iselection ibias iin iperformance ievaluation. iJournal iof iMachine iLearning iResearch i11 i(2010), i2079–2107.
5. Yohan iChon, iNicholas iD. iLane, iFan iLi, iHojung iCha, iand iFeng iZhao. i2012. iAuto- imatically icharacterizing iplaces iwith iopportunistic icrowdsensing iusing ismart- iphones. iIn iProceedings iof ithe iACM iConference ion iUbiquitous iComputing i(UbiComp). iPittsburgh, iPennsylvania, i481–490.
6. Steven iDavis iand iPaul iMermelstein. i1980. iComparison iof iparametric irepresenta- itions ifor imonosyllabic iword irecognition iin icontinuously ispoken isentences. iIEEE iTransactions ion iAcoustics, iSpeech, iand iSignal iProcessing i28, i4 i(1980), i357–366.
7. Gauri iDeshpande iand iBjörn iSchuller. i2020. iAn ioverview ion iaudio, isignal, ispeech, ilanguage iprocessing ifor iCOVID-19. iarXiv ipreprint iarXiv:2005.08579 i(2020).