Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
A machine learning-based approach to determine infection status in recipients of BBV152 (Covaxin) whole-virion inactivated SARS-CoV-2 vaccine for serological surveys

Prateek Singh a,b,1, Rajat Ujjainiya a,b,1, Satyarth Prakash a, Salwa Naushin a,b, Viren Sardana a,b, Nitin Bhatheja a, Ajay Pratap Singh a,b, Joydeb Barman a, Kartik Kumar a, Saurabh Gayali a, Raju Khan b,c, Birendra Singh Rawat d, Karthik Bharadwaj Tallapaka e, Mahesh Anumalla e, Amit Lahiri b,f, Susanta Kar b,f, Vivek Bhosale b,f, Mrigank Sryavastava b,f, Madhav Nilakanth Muggle a,f, C.P. Pandey a,f, Shaziya Khan b,f, Shivani Katiyar b,i, Desh Raj b,f, Sharmeen Ishteyaque b,f, Sonu Khanka b,f, Ankita Rani b,f, Promila b,f, Jyotsna Sharma b,f, Anuradha Seth b,f, Mukul Dutta b,f, Nishant Saurabh f, Murugan Veerapandian b,g, Ganesh Venkatachalam b,g, Deepak Bansal b,h, Dinesh Gupta h, Prakash M. Halami b,i, Muthukumar Serva Peddha b,i, Ravindra P. Veeranna b,i, Anirban Pal b,i, Ranvijay Kumar Singh b,k, Suresh Kumar Anandasadagopan b,i, Parimala Karuppanan l, Syed Nasar Rahman b,l, Gopika Selvakumar l, Subramanian Venkatesan b,i, Malay Kumar Karmakar b,m, Harish Kumar Sardana b,n, Anamika Kothari n, Devendra Singh Parihar b,n, Anupma Thakur b,n, Anas Saini b,n, Naman Gupta b,n, Yogita Singh b,n, Ritu Reddu b,n, Rizul Gautam b,n, Anuj Mishra b,n, Avinash Mishra b,o, Iranna Gogerib,p, Geethavani Rayasamb,d, Yogendra Padwad b,f, Vikram Patial b,r, Vipin Hallan b,r, Damanpreet Singh b,r, Narendra Tirpude b,r, Partha Chakrabarti b,s, Sujay Krishna Maity b,s, Dipyaman Ganguly b,s, Ramakrishna Sistla b,t, Narendar Kumar Balthu t, Kiran Kumar A t, Siva Ranjith t, B. Vijay Kumar 1, Piyush Singh Jamwal t, Anshu Walli u, Sajad Ahmed u, Rekha Chouhan u, Sumit G. Gandhi u,v, Nancy Sharma u,v, Garima Rai u,v, Faisal Irshad b,u, Vijay Lakshmi Jamwal b,u, Masoor Ahmad Paddar b,u, Sameer Ullah Khan b,u, Fayaz Malik b,u, Debashish Ghosh b,v, Ghanshyam Thakkar v, S.K. Barik w,aa, Prabhshau Tripathib,w, Yatendra Kumar Satijaw, Sneha Mohanty b,w, Md. Tauseef Khan b,w, Umakanta Subudhi x,y, Pradip Sen x,y, Rashmi Kumar x,y, Anshu Bhardwaj x,y, Pawan Gupta x,y, Deepak Sharma y, Amit Tuli b,y, Saumya Ray chaudhuri b,y, Srinivasan Krishnamurthi b,y, L. Prakash a, Ch V. Rao b,aa, B.N. Singh b,aa, Arvindkumar Chaurasiya b,ab, Meera Chaurasiyab,ab, Mayuri Bhadangeb,ab, Bhagyashree Lithikarb,ab, Sharada Mohite b,ab, Yogita Patib,ab, Mahesh Kulkarnib,ab, Rakesh Joshib,ab, Vaibhav Pandyab,ab, Sachin Mahajan ab, Amita Patib,ab, Rachel Samson b,ab, Tejas Vare b,ab, Mahesh Darne b,ab, Ashok Girib,ab, Sachin Mahajan ab, Shilpa Paranjapae, G. Narahari Sastry b,ad, Jatin Kalitab,ad, Tridip Phukan b,ad, Prasenjit Manna b,ad, Wahengbam Romib,ad, Pankaj Bharalib,ad, Dibyajyoti Ozb,ad,

Abbreviations: COVID-19, Coronavirus Disease 2019; RT-PCR, Reverse Transcription Polymerase Chain Reaction; SARS-CoV-2, Severe Acute Respiratory Syndrome Coronavirus 2; WHO, World Health Organization; CI, Confidence Interval.

* Corresponding author. CSIR-IGIB, Mathura Road, 110025, Delhi, India.
** Corresponding author. CSIR-IGIB, Mathura Road, 110025, Delhi, India.
E-mail addresses: shantanus@igib.res.in (S. Sengupta), ddash@igib.res.in (D. Dash).

https://doi.org/10.1016/j.compbiomed.2022.105419
Received 24 December 2021; Received in revised form 19 February 2022; Accepted 19 February 2022
Available online 25 April 2022
0010-4825/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
1. Introduction

Mathematical and statistical methods have not only proven helpful to model epidemiological data but also handled the ever-growing host-pathogen data to combat COVID-19 effectively. So far, diverse COVID-19 disease outcome models have been developed using electronic health records, epidemiological and symptoms data [1-3]. Transmission rate and viral load kinetics have been studied using mathematical models on vaccination data [4]. Antibody kinetic analysis found 36%
inactivated whole virion vaccine BBV152 has shown encouraging results in protection against COVID-19 [7] and was approved by WHO on November 3rd, 2021, under the Emergency Use Listing (EUL) category. Recent pilot studies have shown BBV152-Covaxin effectiveness based on neutralization and antibody response against variants of SARS-CoV-2 [8-11]. Recent studies based on RT-PCR have also shown that Covaxin had a protection effectiveness of 47% in previously uninfected individuals, after two doses for symptomatic presentation in health care workers [12]. However, these studies lack a method to detect past undetected infections.

The overarching objective of this study was to determine the effectiveness of BBV152 whole virion vaccines in the general population which requires an accurate estimation of the COVID-19 infection status of the recipients. While the vaccine effectiveness is popularly determined through a test-negative design, it is limited to symptomatic cases, who present for RT-PCR testing and their contacts while asymptomatic infections are largely ignored. In contrast, serology-based assessment of vaccine effectiveness becomes more pertinent in the context of the general population, especially where RT-PCR testing is infrequent and asymptomatic infections are largely ignored. In contrast, serology-based assessment of vaccine effectiveness becomes more pertinent in the context of the general population, especially where RT-PCR testing is infrequent and viral load is insignificant [13,14]. Vaccines targeted specifically to the spike protein (anti-S) do not pose a problem since antibodies against Nucleocapsid proteins (anti-NC) is taken as a marker of infection [15]. However, for whole virion vaccines, anti-NC is induced by the vaccine itself [7] that poses challenges in identifying infection status and hence ascertaining the vaccine protection effectiveness.

To address this gap, we developed a hybrid machine learning approach based on serological indicators to anti-NC and anti-S along with other parameters such as prior history of infection (for Covaxin recipients whose serology history was available), days since last vaccination, gender, age and number of doses taken, as these may have an impact on assessing the infection status of an individual [16,17]. Machine learning (ML) based approaches have shown the ability to integrate multiple parameters to provide a holistic impression of the output variable [18]. ML algorithms are tuned based on the assumptions they follow. Unsupervised ML approaches can identify inherent patterns in the data but are biased towards minor differences in the structure of the data. Supervised ML approaches are biased towards learning the best possible function to approximate the output parameters/labels [19,20]. Hence, integrating the unsupervised and supervised ML algorithms towards ascertaining the infection status of an individual could enable a more generalized assessment of the input parameters.

To develop and validate our method, we used the serosurvey data from the CSIR Cohort, a longitudinal cohort that was developed to assess the disease burden across India and to ascertain the stability of antibodies during post-infection/vaccination [15,21,25]. Population-based cohorts help to determine confident estimates of vaccine effectiveness with a heterogenous accommodation of larger geographical regions. To robustly ascertain the infection status, we took the consensus of two approaches - an unsupervised clustering approach and a supervised SVM-based approach followed by an ensemble model for final infection prediction. We also validated the outcomes of ML models using the surrogate Virus Neutralization Test (sVNT). To the best of our knowledge, this is the first work to predict the infection status and protection effectiveness of Covaxin-vaccinated individuals based on serological analysis and clinical history.

2. Methods

2.1. Data/cohort description

The samples analyzed in this study were from a longitudinal cohort of staff, students and their family members belonging to 43 CSIR laboratories and centers of the Council of Scientific and Industrial Research (CSIR) spread across India (CSIR Cohort; [15]) who had taken one or two doses of Covaxin. The longitudinal cohort study was approved by the Institutional Human Ethics Committee of CSIR-IGIB vide approval CSIR-IGIB/HEC/2019–20. To date, samples have been collected from this cohort in three phases: between June–September 2020 (Phase 1; P1), between January 2021–March 2021 (Phase 2; P2), and May–August 2021 (Phase 3; P3). Phase 3 was incidentally bracketed with the COVID-19 second wave (April 2021–August 2021) in the country and dominated by the Delta variant of SARS-CoV-2 [21].

All subjects participated voluntarily and filled out an online questionnaire form which included information on the date of birth, gender, blood group, type of occupation, comorbidities such as Diabetes, Hypertension, Cardiovascular Diseases, etc, diet preferences, mode of travel, symptomatology, vaccine status, hospitalization (if any), and post-vaccine symptomatology (if any). These forms were then downloaded in MS-Excel data format and merged with registration forms filled at the time of sample collection based on unique IDs.

Blood samples (6 mL) were collected for each subject in an EDTA-coated vacutainer and centrifuged at 1800 g for 15 min at 4°C. Separated plasma was stored at –80°C until used to assess antibodies against recombinant protein representing Nucleocapsid (anti-NC) and Spike (anti-S) antigens of SARS-CoV-2 using Elecsys Anti-SARS-CoV-2 kits (Roche Diagnostics) based on Electro-chemiluminescence Immunoassay (ECLIA) according to manufacturer’s procedure. Individuals with a Cut-off index (COI) value of >1.0 and a value of >0.8 units/millilitre (U/mL) were considered to be positive for anti-NC and anti-S antibodies, respectively. Wherever necessary, samples were appropriately diluted for the anti-S antibody measurements [21].

2.1.1. Input parameters to the ML algorithms

The antibodies to Spike (anti-S) and Nucleocapsid (anti-NC) antigens are the primary serological determinants of infection in an individual as they indicate the presence of antibodies. In our collected samples, both Anti-NC and Anti-S show a bimodal distribution and hence indicate the presence of two subgroups in the data. Since the data had a large variance, it was log-transformed.

To ensure a holistic determination of infection status, we included important covariates as inputs to the ML algorithms along with the serological values. First, the number of doses that a person can be administered was either one or two. Two doses of the vaccine have been reported to elicit higher antibody levels compared to one dose [7,16,17]. Second, the number of days between the last vaccination date and the date of sample collection has been found to be an indicator of the increasing or decreasing trend of Anti-S and Anti-NC levels. A difference of greater than 8 weeks may indicate a decline in antibody levels [17]. Studies have also reported a 3-fold decrease in antibody levels post 6 months of the last vaccine dosage [16]. mRNA vaccines also witness a decrease (half-life of 52 days) in antibody levels after 43 days since last vaccination for all ages [22]. Third, gender (Male/Female) can influence differences in antibody levels [16]. Fourth, the age of the person has been reported as an indicator for antibody levels. Though the effect of age has not been reported as significant for BBV152 vaccines (Covaxin), ChAdOx1-nCOV (Covishield) vaccinated individuals aged greater than 60 years have been reported to have lower antibody formation compared to other age groups [16]. The covariates vary in importance to the prediction process. A PCA biplot was performed to ascertain the contribution of these covariates in explaining the variance of the data.
and their relationship with each other (Supplementary Fig. S1). The inclusion of these covariates as input further reduces bias in the output thereby increasing robustness of the methods.

2.2. Algorithm development

For a supervised and an unsupervised approach, Phase 3 (P3) samples with sero history information in Phase 1 and/or Phase 2 (P1/P2) were included. In P3, people were categorized as “self-reported infected” if they had confirmed RT-PCR results or no RT-PCR but with symptoms (“not confirmed”). Similarly, people without any symptoms or with RT-PCR negative reports were categorized as “self-reported not infected”. Those who were seropositive in P1/P2 (before vaccination was initiated in India), were henceforth considered as infected irrespective of their RT-PCR status or symptomatology. This was an important consideration as P2 negative samples helped us ascertain protection against Delta strain, as that was the period when Delta strain was the predominant variant. Finally, 1063 out of 1823 individuals qualified as input data for downstream analyses (414 infected and 649 not infected samples). The workflow for the pipeline to predict the final infection status is depicted in Fig. 1.

The elements of input data included serological assays - qualitative anti-NC (COI) and quantitative anti-S (U/mL) values of samples along with age, gender, number of vaccination doses and days since the last vaccination. Further, COI, U/mL values and days since the last vaccination were log10 transformed and distribution was calculated using density plots. The supervised approach included SVM-based model development over 100 iterations with the random splitting of 70:30 ratio of training and testing datasets, where final labels were assigned based on the consensus of 100 iterations. Unsupervised clustering itself consisted of two orthogonal approaches, concordance of which was carried forward as cluster 1 and cluster 2. The consensus of the supervised and the unsupervised approaches in addition to P1/P2 seropositive individuals was further used for the final ensemble ML model to predict the full dataset (Fig. 1). Both the approaches are explained in the following sections.

2.2.1. Unsupervised clustering approach

For robustness of the unsupervised analysis, we used two separate approaches namely K-prototype and VarSelLCM. These methods have the best performance benchmark on heterogeneous data for distance-based and model-based clustering methods, respectively, as per a recent comparative study [23]. K-prototype [24] is a distance-based mixed-mode (numerical and categorical) clustering method that uses distance as a measure to define the clusters while VarSelLCM [25] is a mixed-mode model-based method that uses probabilities derived from a model defined using Latent Class Analysis for defining the clusters. The number of clusters best describing the data was first established based on silhouette score (for the K-prototype algorithm) while the VarSelLCM clustering algorithm chooses the best number of clusters automatically. The concordance of cluster assignment between the two methods was ascertained as Cluster 1 and Cluster 2 and discordant labels were labelled ‘Undetermined’. Cluster 1 was predominantly the region that included self-reported infected samples while cluster 2 had a wider dispersion and included self-reported non-infected samples.

2.2.2. Supervised machine learning

Input data features were preprocessed for feature scaling using the standard scaler library in scikit-learn [26] and data was split in a 70:30 ratio for training and testing sets, respectively. Briefly, to ensure that each data point gets a chance to be treated as test data multiple times, 70% (training set) of 1063 entries were trained on the machine learning
model and the rest 30% (319) was used as (blind set) for validation. In every iteration, data splitting from 1063 samples was randomized i.e. no two iterations had identical data as training set or blind set for testing.

The SVM uses a hyperplane with a boundary parameter to separate two classes. It was used for model development as it applies transformation on multi-dimensional data to learn optimal separating parameters for two classes of data. SVM, unlike similar ML algorithms like Logistic Regression, needs to maximize margin for only a subset of points. Therefore, it achieves a better optimal global solution and thus is practically better performing and computationally efficient than most of the algorithms [18]. The hyperparameters for the SVM were chosen based on Grid Search (Supplementary Table S3). The SVM was trained on 70% of training data with a 5-fold cross-validation technique where grid search was used for best parameters selection and validated on a 30% blind dataset. 100 such iterations each with an area under the receiver operating characteristic curve (AUROC) of greater than 0.80 was performed.

The robustness of ML prediction was ensured by iterating the whole pipeline 100 times. A sample was considered infected or not infected only if they were classified as such in at least 75% of the total times predicted, else they were marked as indeterminate.

2.2.3. Ensemble clustering

The outputs of the unsupervised and the ML-based analysis were then compared. We found the concordance between ML-based infected samples with cluster 1 of consensus clustering approach (labelled infected class) and ML-based uninfected samples with cluster 2 of consensus clustering approach (labelled uninfected class). About 1.8% (19/1063), who had a previous positive history of seropositivity but classified as not infected/cluster 2 or indeterminate by both the methods, were reassigned to be positive. These consensus samples were then used to develop an ensemble of 5 ML classifiers (SVM, Logisitic Regression, K-Nearest Neighbors, Random Forest, and Gradient Boost). The hyperparameters for each algorithm were chosen based on Grid Search (Supplementary Table S3). Voting over the outputs of ML algorithms with varying data assumptions provides more specific predictions compared to each individual algorithm. The trained ensemble model was then used to predict the infection status of all the samples.

2.3. Validation of potential asymptomatic samples with neutralizing antibodies

The final model was validated by testing a few samples for their neutralizing activity against Delta infection. For this, a surrogate neutralization was performed against both wild-type and Delta Receptor-Binding Domain (RBD) using GenScript cPass SARS-CoV-2 surrogate Neutralization Antibody detection kit (sVNT) assay (GenScript, USA) as per manufacturer’s instruction [15]. Since the vaccine was developed against the wild-type variant, it is expected that the neutralization will be higher if the neutralizing activity is due to vaccination or infection with wild-type virus or both. However, if a person has been infected with Delta variant, then the neutralizing activity against the Delta RBD will be higher or equal to that of wild-type RBD. Where required, samples were diluted five times. A neutralization of 30% or more in undiluted samples was considered to be positive. To determine the Delta Infection status of the remaining, we first determined the standard error between technical replicates of the samples at various inhibition percentages. Using this, we decided on the Lower bound (LB) of the wild-type inhibition percentage and Upper bound (UB) of the delta-type inhibition percentage as inhibition percentage subtracted and added by twice the average standard error, respectively. Using these criteria, we called any sample with UB of delta > LB of wild-type to be a possible Delta infection.

2.4. Vaccine effectiveness calculation

The unvaccinated group comprised 910 participants who were negative in Phase 2 of the CSIR cohort study. Of these, 567 had become seropositive in Phase 3 of sample collection, while 349 remained uninfected. Of the 164 subjects who took Covaxin from January–May 2021 and were negative in Phase 2, 45 were predicted to be positive by the ML algorithm. Thus, protection was calculated using the following formula [35];

\[(1-RR) \times 100\]

where RR is the relative risk for the vaccinated group to the unvaccinated group.

2.5. Statistical analysis

We used the R software version 3.5.1 for the data curation, management, and clustering analysis. We used the Python library, namely, Scikit-learn (version 0.24.1) for predictive modeling.

3. Results

3.1. Data/cohort details

Blood samples from 1823 Covaxin administered participants belonging to phase 3 (22 May 2021–Aug 09, 2021) of CSIR-cohort were processed for the anti-NC and anti-S antibody assays. Location-wise distribution of individuals has been described in Supplementary Table S1. The details of baseline characteristics are presented in Supplementary Table S2. Out of the 1823 individuals, 772 had taken one dose while 1051 had taken two doses of Covaxin. Of these, 789 and 792 individuals provided samples in P1 and P2, respectively (Fig. 2A).

The preliminary density distribution in terms of anti-NC and anti-S antibodies showed bimodal distribution indicating the presence of two subgroups (Fig. 2B). Further, PCA transformation of the full data, including similar parameters given for supervised/unsupervised approaches (see methods), was analyzed using the top three principal components PC1, PC2, and PC3, which explained 31.1%, 25.1%, and 17.3% of the variance of the data, respectively (Supplementary Fig. S1 (B)). On exploring the PCA biplot (Supplementary Fig. S1(A)) of these variables, anti-NC and anti-S were found to play an important role in explaining the variance of the data. The age of the individual and the days since the last vaccination also contributes to the variance in an orthogonal direction. The number of doses and gender showed limited contribution to the variance. The PCA plot (Fig. 2C) shows the distribution of self-reported infected (red color) and uninfected individuals (grey color). Self-reported infected individuals formed a cluster, while the self-reported negative individuals were spread throughout. It should be noted that although confirmed positive RT-PCR or presence of symptoms most likely indicate infection, a negative PCR test or absence of symptoms does not rule out infection since the RT-PCR test is positive only during a short window of infection [27]. Also, we and others have earlier shown that a large proportion of individuals infected with SARS-CoV-2 are asymptomatic [15,28–30]. This was also observed when the anti-NC and anti-S antibody levels were plotted for self-reported positive and negative subjects (Supplementary Fig. S2). This was irrespective of the vaccine doses, since self-reported positive subjects with one (N = 176) and two doses (N = 199) had similar levels of antibodies forming a single cluster, while those of self-reported negative with one (N = 596) or two doses (N = 852) showed two distinct clusters and had large contours of distribution (Fig. 2D).

The observations are suggestive of a probable undiagnosed infection among the self-reported uninfected individuals. To ascertain the extent of infection in these individuals, we developed a two-step workflow consisting of unsupervised clustering and supervised SVM-based ML
model in parallel, followed by an ensemble model whose inputs consisted of consensus classifications obtained from the two methods in Step 1 (Fig. 1).

3.1.1. Unsupervised clustering identifies two clusters

1063 participants, whose prior serology status at phase 1 or phase 2 was available, were subject to K-prototype and VarSelLCM clustering algorithms. To consider all relevant confounders while creating the clusters, log10 of Antibody to Nucleocapsid, log10 of Antibody to Spike, age, gender, number of doses, and days since last vaccination (log10) for each person were provided as input to the algorithms. Statistical methods in respective clustering methods identified the ideal number of clusters to be two (see methods). Between the two methods, there was a concordance of 96.05% (1021/1063) as shown in the consensus clustering output, while 42 (3.95%) discordant samples were mainly concentrated at the junction of the two clusters (black dots, Fig. 3A). Of the 236 self-reported infected samples, 93.2% (220/236) were found to be in cluster 1, while of the 827 self-reported uninfected, 50.4% (417/827) were in cluster 2.

3.1.2. Supervised ML to predict infection rate

In parallel to the clustering approach, we also developed an SVM-based ML model. For this, as before, we used the 1063 samples of which 414 samples which either had an earlier (P1/P2) seropositive history or self-reported infection status in P3 were considered to be infected, and 649 samples with previous seronegative history and self-reported uninfected status at P3 were considered to be uninfected. SVM-based ML model with 5-fold cross-validation was built on 70% of the randomly selected data (N = 744), while the same model was tested on 30% of the data (N = 319), as mentioned in the methods. If the prediction on blind data was >0.80 AUROC, only then that prediction iteration was considered valid, and this step was repeated for 100 valid iterations. This process reinforced the robustness of the pipeline, where each sample was tested at least 16 times (ranging from 16 to 44).

The average accuracy and AUROC obtained after 100 iterations were 81.3% and 0.88, respectively (Supplementary Fig. S3). The predicted labels over 100 iterations were plotted as sigmoid curves, and samples that were predicted to be positive (531/1063; 49.95%) or negative (493/1063; 46.37%) in at least 75% of the models were labelled infected and uninfected, respectively, while the rest (39/1063; 3.66%) were indeterminate (Fig. 3B).

3.1.3. Ensemble ML model results

We calculated the concordance of ML and clustering approaches with their respective final 1063 labels. Cluster 1 and cluster 2 were enriched in infected and not infected individuals, respectively. We found 526
individuals predicted to be infected by ML and were present in Cluster 1, while 424 individuals were predicted to be uninfected by ML and present in Cluster 2, overall leading to 89.3% concordance. There was one sample which was indeterminate in both cases. As expected, P1/P2 seropositive samples (N = 282) were mostly enriched in the concordant infected panel. The 19 P1/P2 seropositive samples (6.7%) that were assigned negative or indeterminate in the consensus model were reclassified as positive (infected) to build the final model as the seropositivity indicated definite infection. Out of these 19 samples, 4 were classified as concordant “not infected” but were reclassified as infected based on their P1/P2 seropositivity status. Supplementary Fig. S4 shows one to one comparison of three categories of both ML and clustering approaches.

For the final ML ensemble model, we picked concordant (infected, red-colored (N = 545) as well as not infected grey-colored (N = 420)) samples. An ensemble model based on 965 concordant samples was developed using 5 ML algorithms, namely SVM, RF, Gradient Boost, KNN, and Logistic Regression. Using this final model, we assigned infection status to all the 1823 subjects who had taken Covaxin (Fig. 3C). The ensemble model was able to correctly recognize as infected 95% (356/375) of the self-reported infected and 99% (279/282) of the P1/P2 seropositive individuals. Further, the model also assigned 50% (724/1448) of the self-reported negative samples as infected. The prediction on blind 858 individuals stratified by self-reported infection status and the number of doses is represented in Supplementary Fig. S5.

3.2. Validation

The final predictions of the ensemble model were validated using surrogate virus neutralization (sVNT) assays as described in the methods section. All the vaccines have been developed against the wild-type SARS-CoV-2 strain. Hence, it is expected that neutralization against wild-type RBD will be higher if it is due to vaccination or prior infection with the wild-type variant. However, if a person has been infected with
the Delta variant which was predominant (75–80% of all infections [31, 32]) during this period across the country, then the neutralization against Delta RBD will be equal to or greater than wild type RBD. We found that 64.1% (100/156) individuals who self-reported uninfected but were predicted to be infected by the ensemble model, neutralized Delta RBD greater than or equal to that of wild type, suggesting that these individuals were probably infected by Delta variant although they were asymptomatic (Supplementary Fig. S6). Of these, 71.8% (28/39) who were seronegative in P2 i.e., before the second wave and self-reported negative, were found to have higher Delta neutralization (Fig. 3D), which was similar to the frequency of Delta infection across the country.

3.3. Vaccine effectiveness results

Based on the ML outcomes, the vaccine effectiveness was calculated against the unvaccinated group. Amongst the samples collected in P3 of CSIR-Cohort, 34% were unvaccinated of which 567/916 (61.9%) subjects, who were uninfected in P2, were found to be infected in P3 (based on sero values). Thus, the protection effectiveness of Covaxin was calculated to be 55.67% (95% CI 42.9% – 65.6%) after two doses of vaccine.

4. Discussion

It is difficult to distinguish the systemic immune response elicited by the whole virion vaccine from the pathogen infection response of the host. We identified two subgroups of people agnostic to their self-reported infection status from the serosurvey data of Covaxin administered individuals based on the analysis of anti-NC and anti-S antibody response. One subgroup was highly enriched in self-reported positive individuals, while the self-reported negative individuals were distributed across the two subgroups. We hypothesized that the self-reported negative individuals who are in the same group as that of self-reported positive individuals may have had silent undetected (asymptomatic) infections.

To independently determine the infection status of RT-PCR negative individuals, we obtained the seropositivity status in the previous phases of the serosurvey, along with the self-reported infection status. These datasets were run through supervised and unsupervised machine learning approaches. A consensus was built on the infection status for 90.8% (965 out of 1063) individuals. To further validate the infection status proposed by our pipeline, sVNT assays were performed on samples from Phase 3 individuals, which lead to identification of Delta infected individuals who were self-reported as uninfected. The Delta infected samples were highly correspondent with the ML predicted infected people, thereby reinforcing our hypothesis. Therefore, in this study, we provide a machine learning method for objectively annotating the infection status of individuals. This method can be invaluable in predicting the infection status of participants administered Covaxin.

This study fills a gap in the field, however it comes with certain limitations. First, the self-reported status is questionnaire-based, which comes with a level of inconsistency. Second, uninfected individuals might get infected at any time between the questionnaire filling and sample collection. Third, the samples come from employees of different CSIR institutes and their relatives, which might not be the real representation of the overall country’s population, especially in rural areas. Fourth, ML methods can only learn a representation of the available data. Therefore, despite best efforts to reduce bias, the models may have difficulty generalizing to populations that are under-represented in the survey. The geographical locations covered by our study include CSIR labs located throughout the country [15], which complements the predominantly rural locale by the ICMR study [33]. Further, a recent ICMR pilot study with 114 participants showed that individuals with infection and 1 dose of Covaxin developed similar antibody levels to that of individuals who are infection-naïve and received two doses of Covaxin [34].

Finally, we were able to calculate the vaccine protection effectiveness (PE) using our method. We found that Covaxin has a PE of 55.67% (95% CI 42.9%–65.6%) for fully vaccinated subjects. This was similar to the recently published real-world effectiveness of 47% (95% CI 29%–61%) after two doses in previously uninfected individuals who had a symptomatic presentation [12]. However, phase 3 trial results of the vaccine reported 63.6% (95% CI 29.0%–82.4%) PE against asymptomatic infection after two doses of vaccine [7]. This proves the importance of our work and how the ML-based approach could be utilized to study real-world effectiveness, even in asymptomatic population-based on serological methods. Our study was able to address an important gap in the literature for calculating vaccine effectiveness in the case of whole virion vaccines from serology-based surveys.

5. Conclusion

Compared to Spike-inducing vaccines where an increase in the Nucleocapsid antibody levels can be taken as an indicator for infection, whole-virion vaccines such as BBV152, induce both spike and nucleocapsid antibodies. This makes serological determination of infection status in the general population non-trivial. As RT-PCR tests may miss significant proportion of asymptomatic infections, determination of infection status using serological indicators (Anti-S and Anti-NC) becomes important for studies on vaccine effectiveness.

We developed a hybrid ML-based approach, integrating unsupervised and supervised learning to ascertain the infection status. Further, we built an ensemble model (five ML algorithms) that combines serological indicators with important demographic parameters to predict prior infection status in partially and fully vaccinated Covaxin recipients with significant accuracy. Using this approach, we could observe a Protection Effectiveness of 55.67% (95% CI 42.9%–65.6%) in fully vaccinated subjects. This work fills an important gap towards identifying the infection status of whole virion vaccine recipients through serological data in conjunction with an ML-based approach.

Role of the funding source

The sponsor of this study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interest

We declare no conflict of interest.

Funding

Council of Scientific and Industrial Research, India (CSIR).

Acknowledgment

SSG would like to acknowledge CSIR grant MLP 2007 and MLP 1811 for this work. Salwa Naushin and Rajat Ujjainiya would like to thank CSIR for their fellowships. Prateek Singh would like to thank DBT-BINC for their fellowships. Satyarth Prakash would like to acknowledge CSIR grant MLP-1811(CSIR-IGIB) for the salary support. Nitiin Bhatheja would like to acknowledge DST grant GAP-0192 (CSIR-IGIB) for the salary support. Ajay Pratap Singh would like to acknowledge CSIR grant MLP-1811(CSIR-IGIB) for the salary support. Ayushi Singh, Md. Abubakar Sadique from AMPRI. V Santosh Kumar from CCMB. Dr. Shalini Gupta, Shraddha, and Nandpal Yadav from CDRI. PK Khanna, Vipul Sharma, Hansraj Choudhary, Nalini Pareek, Maninder

P. Singh et al.
of a randomised, double-blind, controlled phase 3 trial, Lancet 398 (2021) 2173–2184, https://doi.org/10.1016/S0140-6736(21)00200-6.

[8] G.P. Sankpal, P.D. Yadav, R. Ella, G.R. Deshpande, S. Gupta, K. Gupta, K. Madureva, P. Abraham, S. Panda, B. Bhargava, Inactivated COVID-19 vaccine BBV152/COVAXIN effectively neutralizes recently emerged B.1.1.7 variant of SARS-CoV-2, J. Trav. Med. 28 (2021), https://doi.org/10.1177/1365838321100516.

G. Sankpal, P.D. Yadav, R. Ella, G.R. Deshpande, S. Gupta, K. Gupta, K. Madureva, P. Abraham, S. Panda, B. Bhargava, Inactivated COVID-19 vaccine BBV152/COVAXIN effectively neutralizes recently emerged B.1.1.7 variant of SARS-CoV-2, J. Trav. Med. 28 (2021), https://doi.org/10.1177/1365838321100516.

P. Yadav, G.N. Sankpal, R. Ella, R.R. Sahay, D.A. Nyayanit, D. Desai, P. Abraham, S. Panda, B. Bhargava, Neutralization of Beta and Delta variant with sera of COVID-19 recovered cases and recipients of Covaxin an inactivated COVID-19 vaccine, J. Travel (2021), https://academic.oup.com/jtm/article-abstract/28/3/205/taab077/6277044.

P. Yadav, G.N. Sankpal, R. Ella, R.R. Sahay, D.A. Nyayanit, D. Desai, P. Abraham, S. Panda, B. Bhargava, Neutralization of Beta and Delta variant with sera of COVID-19 recovered cases and recipients of inactivated COVID-19 vaccine BBV152/COVAXIN, J. Trav. Med. 28 (2021), https://doi.org/10.1177/1365838321100516.

Appendix A. Supplementary data
from SERC. Rashmi Arya from URDIP.

Appendix A. Supplementary data

References

[1] H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.H. Strasser, J.G. Klann, P. Naseri, K.B. Wagholikar, S.N. Murphy, H. Estiri, Z.
Y. Bai, L. Yao, T. Wei, B. Flach, J.E. Ledingworth, J.R. Muscola, B.S. Graham, B.C. Lin, S. O’Dell, S.D. Schmidt, A. T. Widge, V.-V. Edara, E.J. Anderson, L. Lai, K. Floyd, N.G. Rospuy, V. Zarnitsyna, F.C. Roberts, M. Makhene, W. Buchanam, C.J. Luke, J.H. Beigel, L. Jackson, K. Jackson, C. Chen, S. Nadeau, M. Vann, P. Voinov, N. Xie, C. Roemer, T. Stadler, CoV-2 spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants, Bioinformatics (2021), https://doi.org/10.1093/bioinformatics/bzt856.

M.V. Murhekar, T. Bhatnagar, S. Selvaraju, K. Rade, V. Saravanakumar, J. W. Vivian Thangaraj, M.S. Kumar, N. Shah, R. Sabarinath, A. Turuk, P.K. Anand, S. Anantha, R. Balachandran, S.D. Bangar, A.K. Bansal, J. Bhat, D. Chakraborty, C. Rangaraju, V. Chopra, D. Das, A.K. Deb, K.R. Devi, G.R. Dwivedi, S.M. Salim Khan, L. Haq, M.S. Kumar, A. Laxmaiah, Madhuka, A. Mhapatra, A. Mitra, A. R. Nirmala, A. Pagdhune, M.A. Qurieshi, T. Ramarao, S. Sahay, Y.K. Sharma, M.B. Shrinivasa, V.K. Shukla, P.K. Singh, A. Viramgami, V.C. Wilson, R. Yadav, C. Girish Kumar, H.E. Luke, U.D. Ranganathan, S. Babu, K. Sekar, P.D. Yadav, G. N. Sapkal, A. Das, P. Das, S. Dutta, R. Hemalatha, A. Kumar, N. Narain, S. Narasimhaiah, S. Panda, S. Patil, S. Patil, K. Sarkar, S. Singh, K. Tripathy, G.S. Toteja, G.R. Babu, S. Kant, J.P. Mulyil, R.M. Pandey, S. Sarkar, S.K. Singh, S. Zodpey, R.R. Gangakhedkar, D.C. Reddy, B. Bhargava, Prevalence of SARS-CoV-2 infection in India: findings from the national serosurvey, May–June 2020, Indian J. Med. Res. 152 (2020) 48–60, https://doi.org/10.4103/ijmr.IJMR_3290_20.

N.P. Kumar, C. Padmapriyadarsini, K.R. Uma Devi, V.V. Banurekha, A. Nancy, C. P. Girish Kumar, M.V. Murhekar, N. Gupta, S. Panda, S. Babu, B. Bhargava, Antibody responses to the BBV152 vaccine in individuals previously infected with SARS-CoV-2: a pilot study, Indian J. Med. Res. 153 (2021) 671–676, https://doi.org/10.4103/ijmr.IJMR_2066_21.

Rajat Ujjainiya, et al., High failure rate of ChAdOx1-nCoV19 immunization against asymptomatic infection in healthcare workers during a Delta variant surge, Nat. Commun. (2022), https://doi.org/10.1038/s41467-022-29404-3.