Clasificación de variantes patogénicas para Diagnóstico Genético

Natalia Castejón Fernández

Esta documentación será revisada y modificada de cara a la versión final, pública, del repositorio una vez los modelos hayan sido correctamente entrenados y demostrada su eficacia.

INTRODUCCIÓN

El diagnóstico genético se define como el análisis del material genético (ADN o ARN), a partir de muestra biológica, con el fin de detectar variantes de la secuencia referencia que puedan estar asociadas a enfermedad (denominadas variantes patogénicas). Normalmente estas variantes presentan frecuencias inferiores al 1% en la población, aunque existen excepciones más extendidas. Las variantes que presentan frecuencias superiores a estas cifras se denominan polimorfismos.

Las metodologías empleadas en el análisis genético son muy variadas. El análisis directo tiene por objetivo identificar o descartar una mutación patogénica en un determinado gen. Sin embargo, tan solo con base en la clínica presentada por el paciente no es fácil determinar qué gen en concreto es el que debe ser analizado. En muchos casos los pacientes no presentan todos los síntomas asociados a una enfermedad y varias enfermedades compartes sintomatologías, por lo que finalmente se deben analizar varios genes para dar con la variante causal. Por otro lado, la existencia de polimorfismos con frecuencias inferiores al 1% y los fallos provocados por las técnicas de secuenciación empleadas, provocan que el analista biológico acabe enfrentándose al análisis individual de varios cientos de variantes candidatas.

Para poder determinar, de entre todas, la variante causal el analista biológico debe considerar cada variante candidata y estudiarla, buscando bibliografía que relacione dicha variante con la clínica que presenta el paciente. Para ayudar al proceso de diagnóstico la comunidad científica ha generado una serie de medidas predictoras de patogenicidad, que permiten, en teoría, priorizar las variantes causales de tal forma que el analista comience analizando aquellas variantes con mayor probabilidad de ser patogénicas. Aun así, las variantes finales que el analista debe comprobar siguen siendo numerosas, lo que convierte el análisis genético en una técnica costosa en tiempo y dinero, que no está al alcance de todo el mundo, amén de que, en muchos casos, el tiempo es crucial debido a que el paciente se encuentra en espera de tratamiento.

Además, desgraciadamente y a pesar de la ayuda que suponen estos predictores, éstos son, en muchos casos, contradictorios ya que se basan en la información obtenida a partir de distintas bases de datos, algunas no correctamente curadas o desactualizadas. Además, los propios valores de los predictores pueden no estar actualizados en la base de datos consultada, por lo que resulta imprescindible valorar varios predictores a la vez. Todo esto complica mucho la labor del analista biológico, en muchos casos o familiarizado con las tareas estadísticas, por lo que se recurre a la labor de los bioinformáticos, que deben priorizar la lista de variantes candidatas de forma que el hallazgo de la variante causal se produzca en el menor tiempo posible.

Por otro lado, en muchos casos las variantes causales aún no han sido descritas, por lo que el analista no puede determinar con certeza la causa de la enfermedad y debe proponer una serie de variantes candidatas que serán estudiadas en ensayos funcionales que permitirán determinar la posible patología asociada a la tenencia de dicha variante. Sin embargo, estos ensayos no solo son muy costosos económicamente, sino que suponen una inversión de tiempo muy levada, postergando el diagnóstico final del paciente durante varios años. Por ello es esencial que el número de variantes candidatas que se destinan a ensayos funcionales sea reducido y que, entre las candidatas derivadas a estos ensayos, se encuentre la variante causal, de forma que la inversión en tiempo y dinero sea finalmente rentabilizada mediante la obtención de un diagnóstico final.

Así, la reducción del número de variantes candidatas por medio de sistemas de clasificación y priorización de la patogenicidad de las mismas, ayudaría a continuar reduciendo el tiempo de diagnóstico, a la par que aumentaría el éxito diagnóstico en aquellos casos en los que aún no se disponga de suficiente bibliografía relacionada, al permitir presentar una variante candidata real con mayor probabilidad de ser la causante.

OBJETIVOS

El presente trabajo pretende presentar un prototipo de clasificación de variantes patogénicas en base a distintos modelos de *machine learning*, que ayude a clasificar las variantes candidatas y reducir el número de variantes a estudiar por los analistas.

Además, este trabajo pretende asentar las bases necesarias para la continuación del desarrollo del modelo que mejores métricas presente en la clasificación.

MATERIALES Y METODOS

Muestras

Para realizar nuestro prototipo partimos de muestras biológicas de 80 pacientes, previamente diagnosticados en nuestro laboratorio como portadores de variantes causales descritas como patogénicas. Todos los patrones de herencia mendeliana quedaron representados en la muestra.

Debido a los contratos de confidencialidad firmados con los pacientes no es posible aportar los datos individualizados de los mismos, por lo que tan solo expondremos la forma en que se obtuvo la tabla final *AllCausalVariats.tx*', que contiene 110 variantes causales patogénicas y 686 664 variantes polimórficas.

Tras el procesamiento y secuenciación de las muestras biológicas de cada paciente, los archivos resultantes fueron anotados según se indica en el archivo *README.md* disponible en el repositorio GitHub. Posteriormente las variantes de cada paciente fueron manualmente anotadas como causante o no causante en un csv. Finalmente, empleando el script *TrainDataFormat.py* se generó la tabla final compartida por Drive.

Por otro lado, en el Drive pueden encontrarse dos carpetas, una llamada /parapredecir y otra denominada /*Validacion*. Dentro de estas carpetas encontramos varios archivos que simulan pertenecer a pacientes. Estos archivos han sido generados a partir de archivos obtenidos de la base de datos *1000GenomesDB* y se aportan a modo de testeo de los scripts del repositorio. Las pruebas de los modelos han sido realizadas con archivos procedentes de pacientes reales que no pueden ser compartidos por la razón anteriormente expuesta. Las variantes causales de estos archivos han sido aleatoriamente escogidas, por lo que las métricas aportadas por el uso de los modelos indican, con mucha probabilidad, malas predicciones. El script *generarDatosparapredecir.py*, disponible en la carpeta /*parapredecir* toma como input uno de los archivos de pacientes y elimina la columna *"causal"*, de forma que estos archivos simulan los archivos que posteriormente se emplearán para predecir las variantes causales en pacientes reales. La carpeta Validación muestra un ejemplo de la estructura de los archivos que se han empleado como set de validación de los modelos entrenados.

Estos archivos no fueron empleados durante el entrenamiento, testeo y validación de los algoritmos presentados y se aportan tan solo a modo de ejemplo para facilitar de la tarea de corrección del presente trabajo.

Campos

Tras la anotación de los archivos se obtienen las siguientes columnas. Para simplificar la extensión de la presente memoria tan solo se explican aquellos campos empleados en la construcción de los modelos.

Chr	**Func.refGene**	**allele_coverage**
Start	**Func.ensGene**	**allele_ratio**
Ref	**Gene.refGene**	**function**
Alt_Annovar	**Gene.ensGene**	**protein**
Alt_IR	**ExonicFunc.refGene**	**coding**
Avsnp147	**ExonicFunc.ensGene**	**Grantham**
Genotype	**MutationTaster_score**	**5000Exomes**
Maf	**MutationTaster_pred**	**FATHMM**
Gene	**PROVEAN_score**	**Clinvar**
Causal	**PROVEAN_pred**	**Cosmic**
Transcript	**CADD_raw**	**Go**
GeneDetail.refGene	**CADD_phred**	**Omim**
GeneDetail.ensGene	**phyloP20way_mammalian**	**Pfm**
AAChange.refGene	**Siphy_29way_logOdds**	**Phylop**
AAChange.ensGene	**FATHMM_score**	**gnomAD_genome_ALL**
SIFT_score	**FATHMM_pred**	**1000g2015aug_all**
SIFT_pred	**CLNALLELEID**	**1000G_ALL**
Ljb23_sift	**CLNDN**	**ExAC_ALL**
Shift	**CLNDISDB**	**AF**
Polyphen2_HDIV_score	**CLNREVSTAT**	**AF_male**
Polyphen2_HDIV_pred	**CLNSIG**	**AF_female**
Polyphen2_HVAR_score	**ncbiRefSeqHgmd**	**PopFreqMax**
Polyphen2_HVAR_pred	**cytoBand**	**gnomAD_HommozCounts**
Polyphen		

- SIFT_score, Shift, ljb23_sift

Este predictor evalúa la tolerancia de la proteína a los cambios de aminoácido con respecto a su función. Valores por debajo de 0.5 indican que la variación es probablemente patogénica. Cuanto menor sea el valor, mayor probabilidad de patogenicidad. El rango es de 0 a 1

- Polyphen2_HDIV_score, Polyphen2_HVAR_score, Polyphen

Este predictor evalúa el posible impacto que produce la sustitución de un aminoácido sobre la función y estructura de una proteína en base a consideraciones físicas e interferencias con los ligandos. Cuanto mayor sea el valor, (rango 0 a 1), mayor es la patogenicidad producida por el cambio.

- Func.refGene, Func.ensGene

Estos dos campos indican si la variante afecta regiones exónicas, intrónicas, intergénicas o regiones 5’ y 3’ UTR. La localización de la variante con respecto a las

regiones codificantes es un parámetro a tener en cuenta a la hora de evaluar la patogenicidad de la misma.

VALOR	EXPLICACIÓN
EXONIC	variant overlaps a coding
SPLICING	variant is within 2-bp of a splicing junction (use -splicing_threshold to change this)
NCRNA	variant overlaps a transcript without coding annotation in the gene definition (see Notes below for more explanation)
UTR5	variant overlaps a 5' untranslated region
UTR3	variant overlaps a 3' untranslated region
INTRONIC	variant overlaps an intron
UPSTREAM	variant overlaps 1-kb region upstream of transcription start site
DOWNSTREAM	variant overlaps 1-kb region downtream of transcription end site (use -neargene to change this)
INTERGENIC	variant is in intergenic region

Las categorías se ordenan según la propia base de datos como:

Exonic = **Splicing** > **ncRNA** > **UTR5/UTR3** > **intron** > **upstream/downstream** > **intergenic**

- ExonicFunc.refGene , Exonic.Func.ensGene

Recoge la consecuencia provocada por la variante dentro de la región exónica (región codificante).

VALOR	EXPLICACIÓN
VALOR	Explicación
FRAMESHIFT INSERTION	An insertion of one or more nucleotides that cause frameshift changes in protein coding sequence
FRAMESHIFT DELETION	A deletion of one or more nucleotides that cause frameshift changes in protein coding sequence
FRAMESHIFT BLOCK SUBSTITUTION	A block substitution of one or more nucleotides that cause frameshift changes in protein coding sequence
STOPGAIN	A nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate creation of stop codon at the variant site. For frameshift mutations, the creation of stop codon downstream of the variant will not be counted as "stopgain"!
STOPLOSS	A nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that lead to the immediate elimination of stop codon at the variant site

NONFRAMESHIFT INSERTION	An insertion of 3 or multiples of 3 nucleotides that do not cause frameshift changes in protein coding sequence
NONFRAMESHIFT DELETION	A deletion of 3 or mutliples of 3 nucleotides that do not cause frameshift changes in protein coding sequence
NONFRAMESHIFT BLOCK SUBSTITUTION	A block substitution of one or more nucleotides that do not cause frameshift changes in protein coding sequence
NONSYNONYMOUS SNV	A single nucleotide change that cause an amino acid change
SYNONYMOUS SNV	A single nucleotide change that does not cause an amino acid change
UNKNOWN	Unknown function (due to various errors in the gene structure definition in the database file)

- CLNSIG, clinvar

Indica, la relevancia clínica recogida sobre la variante, si existe la bibliografia.

VALOR	EXPLICACIÓN
VALOR	EXPLICACIÓN
CLINICAL SIGNIFICANCE VALUE	Guidance for use in ClinVar SCV records
BENIGN	Asrecommended by ACMG/AMPfor variants interpreted for Mendelian disorders.
LIKELY BENIGN	Asrecommended by ACMG/AMP (https://www.ncbi.nlm.nih.gov/pubmed/25741868) for variants interpreted for Mendelian disorders.
UNCERTAIN SIGNIFICANCE	Asrecommended by ACMG/AMPfor variants interpreted for Mendelian disorders.
LIKELY PATHOGENIC	Asrecommended by ACMG/AMPfor variants interpreted for Mendelian disorders.
PATHOGENIC	Asrecommended by ACMG/AMPfor variants interpreted for Mendelian disorders. Variants that have low penetrance may be submitted as "Pathogenic"
DRUG RESPONSE	A general term for a variant that affects a drug response, not a disease. We anticipate adding more specific drug response terms based on arecommendation by CPIC. https://www.pharmgkb.org/page/cpicTermProject
ASSOCIATION	For variants identified in a GWAS study and further interpreted for their clinical significance.
RISK FACTOR	For variants that are interpreted not to cause a disorder but to increase the risk.
PROTECTIVE	For variants that decrease the risk of a disorder, including infections.
AFFECTS	For variants that cause a non-disease phenotype, such as lactose intolerance.

CONFLICTING DATA FROM SUBMITTERS	Only for submissions from a consortium, where groups within the consortium have conflicting intepretations of a variant but provide a single submission to ClinVar.
OTHER	If ClinVar does not have the appropriate term for your submission, we ask that you submit "other" as clinical significance and contact us to discuss if there are other terms we should add.

- MutationTaster_score

En este caso se evalúa la frecuencia de la variante y su presencia en cada uno de los estados genotípicos. Variantes que generen codones de stop se asignan automáticamente a efectos patogénicos. Por último, se realiza un cálculo basado en estadísticas de Bayes. Valores para predecir la patogenicidad de la variación.

- PROVEAN_score

Este predictor indica cuando un cambio proteico afecta la función e la proteína. Los rangos van de -14 a 14.

- CADD_raw, CADD_phred

Permite evaluar cuán deletéreas son las inserciones y delecciones de un único nucleótido.

- Siphy_29way_logOdds

Este predictor en escala logarítmica emplea información de la base de datos dbSNP para indicar cuán conservada es una posición genómica. Cuanto mayor sea el valor de este predictor, más conservada es la posición evolutivamente hablando y, por tanto, más deletérea resulta su variación en términos de probabilidad.

- FATHMM_score, FATHMM_pred, FATHMM

Otro predictor que indica la patogenicidad debida al cambio. En este caso indica la significancia de la consideración de que el cambio es patogénico frente a que sea una variante tolerante.

- Phylop, phyloP20way_mammalian

Indica la conservación evolutiva de la posición genómica. El rango va de -20 a 30, siendo os valores positivos indicativo de menor evolución y, por tanto, de sitios de gran conservación cuya variación generaría patogénia.

- 5000Exomes, Maf, gnomAD_genome_ALL, 1000g2015aug_all, 1000G_ALL, ExAC_ALL, AF, PopFreqMax

Frecuencia poblacional del alelo alternativo (la variante).

- Function

Esta columna indica la afección que la función proteica podría tener debido a la variante. Las categorías son: stoploss, nonsense, missense, frame shift insertion, frame shift deletion, frame shift block substitution, non frame shift insertion, non frame shift deletion, non frame shift, synonymous, unknown.

- Grantham

Indica la distancia biológica existente entre 2 aminoacidos (en este caso, el aminoacido referencia y el que se traduce debido a la variante) en términos evolutivos. Para ellos considera 3 aspectos: composición, polaridad y volumen molecular de los mismos. El rango va de 5 a 215. A mayor distancia, mayor es el daño generad por la sustitución.

Formato

Debido a que cada columna procede de una base de datos diferente, los datos presentan distinto forma, separador y tabulado. Por ello el primer paso consiste en homogeneizar el formato de estos. Además, en algunos campos se encuentra más de un dato, por lo que se programa una función que además de separar dichos datos, escoja en cada caso el valor que represente la mayor probabilidad de patogenicidad. Por su parte, las columnas cuyos rangos son diferentes al rango 0-1 se estandarizan, y los campos que representan funciones se ordenan de menor a mayor probabilidad de patogenicidad y se categorizan.

Datos faltantes

Muchos de los datos presentan valores faltantes en algunos de sus campos. Es por ello por lo que para rellenar esos valores se decidió calcular la correlación entre distintos campos para rellenar valores faltantes de unas a partir de las otras.

Así, se generan nuevas variables y se reduce dimensionalidad.

Debemos señalar que en casi todos los casos se pudo emplear una regresión lineal para predecir el valor continuo de los predictores numéricos, salvo en uno, donde se empleó una regresión logarítmica transformando una variable en su logaritmo debido a que la relación entre las mismas era monotónica y no lineal. En el caso de predictores categóricos se empleó la regresión logística. Todo esto se realiza en el script "estatstics.py".

Los modelos resultantes se guardan en la carpeta */MLmodel* y son posteriormente empleados en el script de predicción para rellenar los valores faltantes.

Aumento de la proporción de eventos causales

Debido a la naturaleza de los datos, la cantidad de variantes causales es muy baja con respecto a la de eventos no causales. Esto puede generar sesgos en el entrenamiento de los modelos.

Por ello, las variantes causales son retenidas y las no causales son filtradas para eliminar gran parte de ellas en base a criterios como su frecuencia poblacional o la pertenencia a categorías que indican tolerancia. Posteriormente las variantes causales son nuevamente añadidas.

Además, los modelos fueron entrenados con estos datos y con os mismos tras la aplicación de un modelo de "upsampling" basado en SMOTENC, un SMOTE que puede ser aplicado cuando los datos tienen variables categóricas.

Clasificación

Se probaron distintos modelos. El primero, un modelo naive basado en la proporción de aparición de las variantes causales. (en el caso de los datos a los que se aplicó SMOTENC, esta proporción es 0.5, convirtiendo el modelo de Bernoulli en una binomial).

Además, se probaron modelos basados en *Decission Tree* y en *Gradient Boosted*. Finalmente se entrenó una red convolucional sencilla. Todos estos modelos son guardados para su posterior aplicación en la predicción con el script "predict.py".

Debemos indicar que el modelo basado en redes convolucionales genera problemas para ser usado mediante carga con *pickle* debidas a que se ha tenido que definir la medida de *recall*, por lo que la predicción con este modelo se entrega comentada.

La métrica más importante, debido a la naturaleza de los datos es la *recall*, ya que lo más interesante es poder predecir correctamente la variante patogénica. También resulta interesante considerar el número de falsos positivos que arroja cada modelo.

RESULTADOS

Las siguientes tablas muestra los coeficientes de Pearson de las distintas variables. Esta tabla se empleó para decidir qué variables se usaban en los modelos de predicción para el rellenado de valores faltantes.

	Pearson	gnomAD_genom	1000g2015aug_all	1000G_ALL	ExAC_ALL	AF	PopFreqMax	5000Exomes
	maf	0.026	0.029	0.035	0.034	0.028	0.219	0.907
	gnomAD_genome_ALL	x	0.98	0.979	0.981	0.982	0.925	0.054
	1000g2015aug_all	x	x	0.999	0.973	0.975	0.948	0.04
	1000G_ALL	x	x	x	0.972	0.974	0.95	0.04
	ExAC_ALL	x	x	x	x	0.995	0.93	0.06
>0.9	AF	x	x	x	x	x	0.928	0.055
>0.75	PopFreqMax	x	x	x	x	x	x	0.223
<0.75	5000Exomes	x	x	x	x	x	x	x

	Pearson	sift	ljb23_sift	...en2_HDIV	...en2_HVAR	polyphen	...20way_mam	phylop	...tionTaster_	...OVEAN_sco	...CADD_phred	FATHMM	grantham	...THMM_sco	..._29way_logOdds
	SIFT_score	0.867	0.809	-0.577	-0.526	-0.566	-0.29	-0.2	0.291	0.672	-0.574	-0.475	-0.118	0.09	-0.195
	sift	x	0.759	-0.574	-0.517	-0.542	-0.295	-0.171	0.295	0.659	-0.556	-0.477	-0.11	0.084	-0.181
	ljb23_sift	x	x	-0.506	-0.466	-0.496	-0.209	-0.13	0.238	0.609	-0.485	-0.382	-0.112	0.064	-0.165
> + - 0.9	...olyphen2_HDIV_s	x	x	x	0.92	0.926	0.363	0.379	-0.354	-0.58	0.668	0.542	0.111	-0.07	0.37
> + - 0.70	...olyphen2_HVAR_s	x	x	x	x	0.876	0.339	0.405	-0.351	-0.569	0.644	0.534	0.106	-0.07	0.388
> + - 0.6	polyphen	x	x	x	x	x	0.365	0.366	-0.346	-0.56	0.663	0.521	0.109	-0.092	0.362
< + - 0.6	...yloP20way_mamm	x	x	x	x	x	x	0.58	-0.292	-0.243	0.524	0.497	0.068	-0.192	0.344
	phylop	x	x	x	x	x	x	x	-0.376	-0.187	0.572	0.614	0.06	-0.219	0.61
	MutationTaster_sc	x	x	x	x	x	x	x	x	0.261	-0.43	-0.607	-0.06	0.102	-0.348
	PROVEAN_scor	x	x	x	x	x	x	x	x	x	-0.498	-0.487	-0.044	0.09	-0.215
	CADD_phred	x	x	x	x	x	x	x	x	x	x	0.683	0.12	-0.143	0.495
	FATHMM	x	x	x	x	x	x	x	x	x	x	x	0.002	-0.21	0.557
	grantham	x	x	x	x	x	x	x	x	x	x	x	x	-0.025	0
	FATHMM_score	x	x	x	x	x	x	x	x	x	x	x	x	x	-0.138
	SiPhy_29way_logO	x	x	x	x	x	x	x	x	x	x	x	x	x	x

	Pearson	ExonicFunc.refGene	function	clinvar
	Func.refGene	0	0	0.068
	ExonicFunc.refGene	x	0.973	0.093
	function	x	x	0.118
> + - 0.9	clinvar	x	x	x
> + - 0.70				
> + - 0.6				
< + - 0.6				

Respecto a los modelos de clasificación, en todos los casos se vio que el entrenamiento con datos a los que se había aplicado SMOTENC (balanced class) dió mejores resultados que el entrenamiento con set de datos original.

	IMBALANCED CLASS		BALANCED CLASS	
NAIVE (BERNNOUILI)	TP: 18 FP: 0 Recall: 0.79 Accuracy: 0.988 Precision: 0.969	FN: 92 TN: 2206	TP: 1692 FP: 3 Recall: 0.792 Accuracy: 0.895 Precision: 0.996	FN: 514 TN: 2203
DECISSION TREE	TP: 103 FP: 0 Recall: 0.9 Accuracy: 0.966 Precision: 0.878	FN: 7 TN: 2206	TP: 2196 FP: 1 Recall: 0.987 Accuracy: 0.905 Precision: 0.906	FN: 10 TN: 2205
GRADIENT BOOSTED	TP: 104 FP: 0 Recall: 0.9 Accuracy: 0.964	FN: 6 TN: 2206	TP: 2182 FP: 9 Recall: 0.987 Accuracy: 0.948	FN: 24 TN: 2197

	Precision: 0.893		Precision: 0.950	
CONVOLUTIONAL NETWORK	TP: 86	FN: 24	TP: 2169	FN: 37
	FP: 0	TN: 2206	FP: 22	TN: 2184
	Recall:		Recall:	
	Train 0.628. Test 0.897		Train 0.982 Test 0.990	
	Accuracy:		Accuracy:	
	Train 0.989 Test 0.996		Train 0.986 Test 0.991	
	Precision:		Precision:	
	Train 0.737 Test 0.966		Train 0.990 Test 0.988	

Tras evaluar las matrices de contingencia de cada modelo y la *recall*, se determina que el mejor modelo es el entrenado en base a un *Gradient Boosted*.

El modelo basado en redes convolucionales sufre sobreajuste, ya que cuando se emplea para predecir sobre datos no entrenados (resultados no adjuntos) el número de falsos positivos aumenta considerablemente.

DISCUSIÓN

El punto fuerte del pipeline desarrollado consiste en la facilidad que presenta para continuar su uso. Las métricas de cada modelo pueden ser des-comentadas o reconducidas a archivos de texto para su comparativa. Además, el script *main.py* dirige todo el entrenamiento, mientras que el script *Comp_w_Test.py* dirige la validación en muestras y el script *predict.py* permite realizar a predicción de las variantes patogénicas y su escritura en un archivo de texto para posterior evaluación o introducción en el pipeline rutinario de análisis bioinformático.

CONCLUSIONES

El presente trabajo sienta las bases sobre las que continuar trabajando para realizar un buen modelo de clasificación. A pesar de que actualmente el modelo *Gradient Boosted* presenta las mejores métricas y una baja tasa de falsos positivos, es un hecho comprobad que las redes convolucionales funcionan mejor en datos genéticos. Tras un estudio de la situación del modelo se puede determinar que actualmente la red convolucional está sobre ajustada, pero su reajuste futuro podría ser la mejor opción de cara a la clasificación e variantes y consecución óptima del objetivo del trabajo.